

高等学校专业英语教材

统计学专业英语教程

王忠玉 宋要武 编著

電子工業出版社

Publishing House of Electronics Industry

北京 · BEIJING

内 容 简 介

本书内容分为三部分：第一部分是描述统计学，共有 7 个单元，包括统计学初步、单变量数据的描述分析、两个变量数据的描述分析、概率初步、离散概率模型、连续概率模型、抽样分布和中心极限定理；第二部分是推断统计学（数理统计学），共有 4 个单元，包括统计推导初步、一个总体的统计推断、两个总体的统计推断、简单回归、回归的统计分析；第三部分是统计学与数据科学专题，只有 1 个单元。

和同类书籍相比，本书具有如下特点：（1）比较系统地阐述基础统计学的知识，即以初阶统计学的基本内容为主体，又适当地加入并介绍中阶统计学的部分内容；（2）在大多数章后，我们提供课外进一步阅读和学习的补充知识，有统计学家简介等；（3）紧跟当今时代发展，给出“统计学与数据科学”的阅读学习内容。另外，在每一章前面，我们精心选取了一些著名统计学家或教授的名言或警句，同时，特别绘制了有趣的漫画。本书提供部分习题参考答案、教学 PPT、音频资料、部分课文参考译文及其他辅助资料，读者可从华信教育资源网 www.hxedu.com.cn 免费下载，也可扫描二维码获取。

本书适合于各个专业对统计学专业英语感兴趣的大学生，学习双语统计学的统计学专业大学生，希望学习和掌握中阶统计学的相关专业的低年级研究生，以及有关的科研人员等。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目 (CIP) 数据

统计学专业英语教程 / 王忠玉，宋要武编著. —北京：电子工业出版社，2016.8

高等学校专业英语教材

ISBN 978-7-121-28928-6

I. ①统… II. ①王… ②宋… III. ①统计学—英语—高等学校—教材 IV. ①H31

中国版本图书馆 CIP 数据核字 (2016) 第 117241 号

策划编辑：秦淑灵

责任编辑：秦淑灵

印 刷：

装 订：

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：787×1 092 1/16 印张：24.75 字数：824 千字

版 次：2016 年 8 月第 1 版

印 次：2016 年 8 月第 1 次印刷

印 数：3000 册 定价：49.80 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010)88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：(010)88254531。

序 言

当前,就我国大学英语教学的目标或模式而言,通常本科基础英语或传统的综合英语模式,不论是理、工、农林类,还是管理、财经类等大学生都要学习基础外语,最终以通过四级或六级英语考试为评价阶段性英语教学任务的标准。然后,各高校因专业不同而开设各自专业的英语,其目标是为大学生有机会接触用英语讲授的专业讲座和专业课程提供查阅、搜索和研究某个专题文献综述的一个良好开端。这样的教学模式可用图 1 来表示,也就是英语教学的实用目标和专业英语、基础英语三者之间的关系。

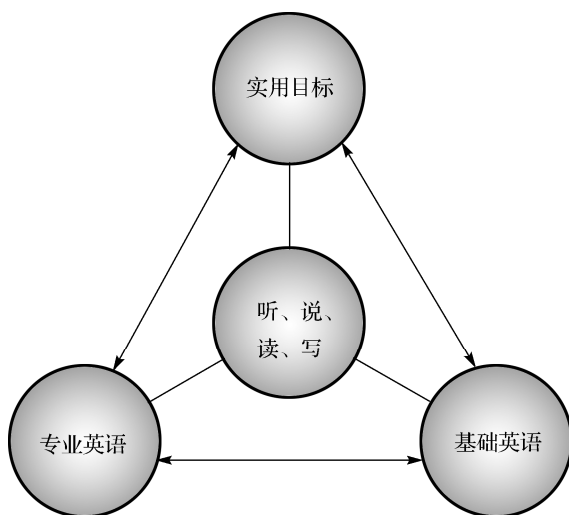


图 1 专业英语、基础英语和实用目标三者关系

2012 年 4 月,上海市大学英语教学指导委员会宣布了以学术英语为导向的指导性文件《上海市大学英语教学参考框架(试行)》。2014 年 4 月,由上海交通大学出版社出版《新核心综合学术英语教程》第四册,从而完成第一套大学生学术英语教材(共 4 册),这表明大学英语的教学方式向以“专门用途英语”为导向的转型之路走出了重要一步。出现这样的变化,可以说反映出一种新的趋势和发展事态,即随着大学教育的日益普及、大学生的外语水平普遍提升,国际交往和交流越来越多,各高校的不同专业大学生接触和联系外国专家、学者的机会也不断增多,试图通过外文直接获取、学习本专业知识的途径非常便利。同时,由于互联网的发展和移动互联网的普及,使得以往制约人们寻找、发现和获取新信息的瓶颈不复存在,进而出现了新的发展趋势。

那么,究竟什么是学术英语呢?学术英语的含义是 *English for Academic Purposes*, 记为 EAP, 一般可以分成两大类:第一类是通用学术英语(*English for General Academic Purposes*, EGAP);第二类是专门学术英语(*English for Specific Academic Purposes*, ESAP)。前者是一个跨越多学科的语言教学,目的是为各专业学生提供所需的通用共性的基本技能,包括学术口语交流能力和学术书面交流能力。具体地说,比如如何听讲座、做笔记、搜索和阅读文献、

撰写课程小论文、参加学术讨论等。后者则是某个特定学科领域（如数学、统计学、生物学、经济学、物理学）的英语教学。

实际上，如果从更广阔的视角来考察，许多高校的博士研究生英语教学早在多年前就已经执行了学术英语的教学。当今，随着英语逐渐成为世界上各个学科交流科研成果、各个学术团体及组织、会议和期刊的通用语言，学术英语迅速扩展到全世界。由此可见，学术英语的目标是培养大学生对本专业文献信息的查阅、搜集、评价、组织及表达的能力。尤其是，开展以问题或项目为指针的教学，使学生具有独立思考、独立学习的研究能力，这是每一名大学生所必备的学术素养（见图 1）。

编写这本书的主要目的是，尝试提供一本针对统计学领域的专业英语，比较系统地阐述基础统计学的知识。作为统计学导论的书籍，本书深入浅出地讲解和阐述什么是统计学，特别是初阶统计学的基本内容。同时，紧跟当今时代发展，整理出“统计学与数据科学”可供选学和课外阅读的单元。

实际上，从数据科学（Data Science）的交叉属性来看，可将数据科学看成计算机科学、数值计算、现代数据分析等的交叉融合而形成的新兴学科，目的是从数据中获得知识，获得有价值的信息，服务于社会。如果从应用视角看，数据科学应具备三个条件：第一个条件是底层构架开发或使用能力，如 Spark, MapReduce, Hadoop 等；第二个条件是程序开发能力；第三个条件是数据建模和解决问题能力。

美国加州大学伯克利分校统计系的郁彬（Bin Yu）教授提出，一个合格的数据科学家应具备的基本素质和技能，可概括为 SDC³：

- Statistics (S) 统计学；
- Domain (science) knowledge (D) 深厚的（科学）知识；
- Computing (C) 计算技术；
- Collaboration (“team work”) (C) 团队的合作能力；
- Communication (to outsiders) (C) 与外界的沟通能力。

并认为

$$\text{Data Science} = \text{SDC}^3$$

美国统计学家吴建民教授（C.F. Jeff Wu）早在 1998 年的一个学术会议上就曾建议：

Statistics → Data Science

Statisticians → Data Scientists

Several good names have been taken up: computer, information science, material science, cognitive science. “Data Science” is likely the remaining good name reserved for us.

为了适应这一技术变革趋势与新兴的社会需求，伊利诺伊大学香槟分校从 2011 年起举办“数据科学暑期研究班”；哥伦比亚大学从 2013 年起开设《应用数据科学》课程，并从 2013 年起开设相关培训项目，从 2014 年起设立硕士学位，2015 年设立博士学位；纽约大学从 2013 年秋季起设立“数据科学”硕士学位。在英国，邓迪大学从 2013 年起设立“数据科学”硕士学位。

特别要提及的是，美国的得克萨斯大学奥斯汀分校（The University of Texas Austin）的自然科学学院（College of Natural Sciences）索性将统计系改名为统计及数据科学系（Department of Statistics and Data Sciences），而其他大学（如美国的西弗吉尼亚大学（West

Virginia University)) 统计系硕士研究生设有数据科学方向 (Master of Data Science)。另外, 斯坦福大学统计学系研究生层面教育也有数据科学方向。由此可见, 统计学是数据科学中最重要的组成部分之一。

作为数据科学三大支柱之一的计算机科学, 迄今为止的发展经历了三个阶段。早期阶段, 让计算机可以工作, 发展重点在于程序语言、编译原理、操作系统以及支撑它们的数学理论; 中期阶段, 让计算机变得有用, 发展重点在于算法和数据结构; 当前阶段, 让计算机具有更多的应用, 发展重点从离散类数学转到概率与统计。

如果从计算机科学处理数据的核心技术看, 机器学习就是当前最核心的技术之一, 而且发展势头非常强劲, 那么就不能不提到统计 (或统计方法)、数据、计算和机器学习这四者的关系, 如图 2 所示。因此, 在“统计学与数据科学”中对机器学习、统计计算等都有所涉及。另外, 有一种观点认为, 机器学习等价于“数据矩阵+统计学+最优化+算法”。由此可见统计学作为数据科学的另一个支柱的重要性。

2014 年 6 月 25 日, 全国科学技术名词审定委员会发布试用 204 条科技新词, 其中包括“大数据”、“云计算”、“物联网”、“三维打印”等 42 条热点名词以及“暗能量”、“宏基因组”等 162 条专业新词。具体地说, 大数据 (big data) 是指具有数量巨大 (无统一标准, 一般认为在 T 级或 P 级以上, 即 10^{12} 或 10^{15} 以上), 类型多样 (既包括数值型数据, 也包括文字、图形、图像、音频、视频等非数值型数据), 处理时效紧, 数据源可靠性保证度低等综合属性的数据集合。再比如, 物联网 (internet of things) 是指综合采用计算机、网络、传感器、控制设备等, 让能够被独立寻址的相关物理对象互联互通, 实现对其识别、监控和管理的智能化网络。(摘自 http://tech.gmw.cn/2014-06/25/content_11727262.htm; http://tech.gmw.cn/2014-06/25/content_11727659.htm。)

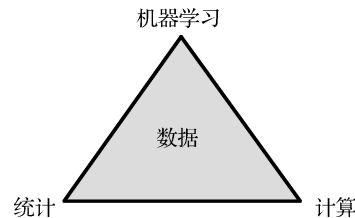


图 2 统计、数据、计算和机器学习四者的关系

本书的“统计学与数据科学”单元就包括了大数据、物联网这样的科技新词。

作者曾经出版过《统计学专业英语 (第 3 版)》(哈尔滨工业大学出版社, 2015 年 4 月), 但是这两本书在内容素材选取、难易程度等方面, 各自有不同的特点, 这两本书的关系可以说是相互补充, 没有替代性。具体而言, 《统计学专业英语 (第 3 版)》是针对学习过统计学或数理统计学的大学生, 提供了不同专题的单元内容以学习和掌握统计学专业英语。而本书则是针对以前没有学习过统计学, 打算了解和掌握利用英语阐明统计学基础知识的大学生, 内容定位为初阶统计学, 内容素材和前者相比, 完全不同。

另外, 本书在内容选取上, 充分兼顾“特性群体的大学生”, 也就是懂一点统计学但又想学习统计学专业英语的各类专业大学生或研究生, 各行各业有这方面需求的工作者。

本书尝试在下述几方面进行探索。

(1) 定位: 这是初阶、中阶统计学领域的专业英语, 目的是使学生初步认识、了解和掌握统计学专业领域的常用术语, 掌握统计学的基本内容, 学会运用基本统计分析方法。

(2) 教学内容: 以单元形式提供相关的统计内容, 给出有关的英文术语及词汇表。另外, 为方便教学, 提供有关用于教学的 PPT 等。

(3) 习题解答: 对某些较难的计算习题, 给出参考答案。另外, 我们为使用本书的教师

提供一些额外的教学资料,可以直接联系编辑或作者(编辑 E-mail: qinshl@phei.com.cn)。

部分习题参考答案、教学 PPT、音频资料、部分课文参考译文及其他辅助资料,可登录华信教育资源网www.hxedu.com.cn免费下载,也可扫描二维码获取。

全书的内容安排以 Unit(单元)为独立形式,前面有一个名言或语录、漫画,然后是单元的详细目录。随后,进入统计知识的阐述和讲解。最后有些单元后面还提供“补充阅读内容”,这包括两类知识:一类是正文内容的补充,另一类是历史人物、重要知识点等。音标所注为英式发音。

与此同时,为了扩展知识面,紧密联系当代统计学的新应用,本书特别编写“统计学和数据科学”的内容,这是一个选学单元。因此,这是一本初阶、中阶统计学专业英语教程,其中有些小节带有星号,表示这样的内容更适合于中阶。

全书整体设计有别于其他教材(不论是国内统计学教材,还是外文统计学教材),目的是试图编写出一本既有专业性,又有趣味性、可读性的教材或参考书。这本书就是在这种理念下构思编写而成的。

另外,本书的附录提供了几个有用的内容,包括统计学领域专业术语的标准翻译,参考了国家标准化委员会发布和出版的关于统计学方面的术语及英文翻译,也就是《GB/T 3358.1-2009 统计学词汇及符号 第 1 部分:一般统计术语与概率的术语》。实际上,如果读者需要学习和了解更多的统计学专业术语,可以参看《GB/T 3358.2-2009 统计学词汇及符号 第 2 部分:应用统计》;《GB/T 3358.3-2009 统计学词汇及符号 第 3 部分:实验设计》。

全书内容安排如下:王忠玉(哈尔滨工业大学经济与管理学院)编写第 1、6、7、8、9、10、12 单元,宋要武(黑龙江科技大学、哈尔滨股权投资协会)编写第 2、3、4、5 单元。哈尔滨理工大学经济学院的张莹老师编写 11 单元和书后附录的整理工作(包括统计学常用术语翻译、附录)。漫画制作人员有黑龙江农垦职业学院李辰光、戈娇老师,哈尔滨师范大学传媒学院曹文龙老师,研究生朱砚。本书采用漫画形式,以此概括或揭示本章内容的某种浓缩特色。这些漫画为书增添了趣味性和可读性。

另外,英国帝国理工大学的数学系学生朱烜繁提供了有益的建议。其他同学,如哈尔滨工业大学的王初旭、邢喆、周子涵、温雅欣、仇派、于娜、陈悦竹、夏晴、牟思涵、郑天慧、范晓菲,另外 2013 级金融学的武杰、满达,2014 级金融学的董赫,黑龙江大学的王天元等,也提供了许多有益的帮助。

书中难免存在纰漏和错误,欢迎广大读者、教师批评指正。E-mail: h20061111@126.com。

王忠玉 宋要武



部分习题参考答案



部分课文参考译文



附录 D-H



音频

目 录

Part I Descriptive Statistics

Unit 1 Statistics	3
1.1 What is Statistics?	4
1.1.1 Meanings of Statistics	4
1.1.2 Definition of Statistics	5
1.1.3 Types of Statistics	6
1.1.4 Applications of Statistics	6
1.2 The language of Statistics	9
1.2.1 Population and Sample	9
1.2.2 Kinds of Variables	11
1.3 Measurability and Variability	14
1.4 Data Collection	16
1.4.1 The Data Collection Process	17
1.4.2 Sampling Frame and Elements	18
1.5* Single-Stage Methods	21
1.5.1 Simple Random Sample	21
1.5.2 Systematic Sample	22
1.6* Multistage Methods	25
1.7* Types of Statistical Study	27
1.8 The Process of a Statistical Study	31
Glossary	34
Reading English Materials	35
Passage 1. What is Statistics?	35
Passage 2. From Data to Foresight	35
Problems	36
Unit 2 Descriptive Analysis of Single-Variable Data	40
2.1 Graphs, Pareto Diagrams, and Stem-and-Leaf Displays	41
2.1.1 Qualitative Data	41
2.1.2 Quantitative Data	43
2.2 Frequency Distributions and Histograms	47
2.2.1 Frequency Distribution	47

2.2.2	Histograms	51
2.2.3	Cumulative Frequency Distribution and Ogives	53
2.3	Measures of Central Tendency	55
2.3.1	Finding the Mean	55
2.3.2	Finding the Median	56
2.3.3	Finding the Mode	57
2.3.4	Finding the Midrange	58
2.4	Measures of Dispersion	60
2.4.1	Sample Standard Deviation	62
2.5	Measures of Position	64
2.5.1	Quartiles	64
2.5.2	Percentiles	64
2.5.3	Other Measures of Position	66
2.6	Interpreting and Understanding Standard Deviation	70
2.6.1	The Empirical Rule and Testing for Normality	70
2.6.2	Chebyshev's Theorem	72
	Glossary	74
	Problems	75
Unit 3	Descriptive Analysis of Bivariate Data	79
3.1	Bivariate Data	80
3.1.1	Two Qualitative Variables	80
3.1.2	One Qualitative and One Quantitative Variable	82
3.1.3	Two Quantitative Variables	83
3.2	Linear Correlation	85
3.2.1	Calculating the Linear Correlation Coefficient, r	86
*3.2.2	Causation and Lurking Variables	89
3.3	Linear Regression	91
3.3.1	Line of Best Fit	92
3.3.2	Making Predictions	97
	Reading English Materials	99
	Passage 1. The First Regression	99
	Passage 2. Simpson's Paradox	99
	Problems	100
Unit 4	Introduction to Probability	104
4.1	Sample Spaces, Events and Sets	105
4.1.1	Introduction	105
4.1.2	Sample Spaces	105
4.1.3	Events	106

4.1.4	Set Theory	108
4.2	Probability Axioms and Simple Counting Problems	109
4.2.1	Probability Axioms and Simple Properties	109
4.2.2	Interpretations of Probability	111
4.2.3	Classical Probability	112
4.2.4	The Multiplication Principle	113
4.3	Permutations and Combinations	115
4.3.1	Introduction	115
4.3.2	Permutations	116
4.3.3	Combinations	118
4.3.4	The Difference Between Permutations and Combinations	120
4.4	Conditional Probability and the Multiplication Rule	122
4.4.1	Conditional Probability	122
4.4.2	The Multiplication Rule	123
4.5	Independent Events, Partitions and Bayes Theorem	124
4.5.1	Independence	124
4.5.2	Partitions	125
4.5.3	Law of Total Probability	126
4.5.4	Bayes Theorem	126
4.5.5	Bayes Theorem for Partitions	127
	Reading English Materials	130
	Passage 1. Probability and Odds	130
	Passage 2. The Relationship between Odds and Probability	130
	Passage 3. How the Odds Change across the Range of the Probability	131
	Problems	132
Unit 5	Discrete Probability Models	134
5.1	Introduction, Mass Functions and Distribution Functions	135
5.1.1	Introduction	135
5.1.2	Probability Mass Functions (PMFs)	136
5.1.3	Cumulative Distribution Functions (CDFs)	137
5.2	Expectation and Variance for Discrete Random Quantities	138
5.2.1	Expectation	138
5.2.2	Variance	139
5.3	Properties of Expectation and Variance	140
5.3.1	Expectation of a Function of a Random Quantity	140
5.3.2	Expectation of a Linear Transformation	140
5.3.3	Expectation of the Sum of Two Random Quantities	141
5.3.4	Expectation of an Independent Product	141

5.3.5	Variance of an Independent Sum	142
5.4	The Binomial Distribution	142
5.4.1	Introduction	142
5.4.2	Bernoulli Random Quantities	143
5.4.3	The Binomial Distribution	143
5.4.4	Expectation and Variance of a Binomial Random Quantity	145
5.5	The Geometric Distribution	146
5.5.1	PMF	146
5.5.2	CDF	147
5.5.3	Useful Series in Probability	148
5.5.4	Expectation and Variance of Geometric Random Quantities	148
5.6	The Poisson Distribution	149
5.6.1	Poisson as the Limit of a Binomial	149
5.6.2	PMF	150
5.6.3	Expectation and Variance of Poisson	151
5.6.4	Sum of Poisson Random Quantities	152
5.6.5	The Poisson Process	152
	Reading English Materials	154
	Passage 1. The Founder of Modern Statistics—Karl Pearson	154
	Passage 2. The Relations of Several Discrete Probability Models	154
	Problems	155
Unit 6	Discrete Probability Models	158
6.1	Introduction, PDF and CDF	159
6.1.1	Introduction	159
6.1.2	The Probability Density Function	159
6.1.3	The Distribution Function	160
6.1.4	Median and Quartiles	161
6.2	Properties of Continuous Random Quantities	161
6.2.1	Expectation and variance of continuous random quantities	161
6.2.2	PDF and CDF of a Linear Transformation	162
6.3	The Uniform Distribution	163
6.4	The Exponential Distribution	165
6.4.1	Definition and Properties	165
6.4.2	Relationship with the Poisson Process	166
6.4.3	The Memoryless Property	167
6.5	The Normal Distribution	168
6.5.1	Definition	168

6.5.2	Properties	168
6.6	The Standard Normal Distribution	169
6.6.1	Properties of the Standard Normal Distribution	170
6.6.2	Finding Area to The Right of $z = 0$	171
6.6.3	Finding Area in The Right Tail of a Normal Curve	171
6.6.4	Finding Area to the Left of a Positive z Value	172
6.6.5	Finding Area from a Negative z to $z = 0$	172
6.6.6	Finding Area in the Left Tail of a Normal Curve	172
6.6.7	Finding Area from A Negative z to a Positive z	172
6.6.8	Finding Area Between two z Values of the Same Sign	173
6.6.9	Finding z -Scores Associated with a Percentile	173
6.6.10	Finding z -scores that Bound an Area	174
6.7	Applications of Normal Distributions	175
6.7.1	Probabilities and Normal Curves	175
6.7.2	Using the Normal Curve and z	176
6.8	Specific z -score	178
6.8.1	Visual Interpretation of $z(a)$	179
6.8.2	Determining Corresponding z Values for $z(a)$	179
6.8.3	Determining z -scores for Bounded Areas	180
6.9	Normal Approximation of Binomial and Poisson	181
6.9.1	Normal Approximation of the Binomial	181
6.9.2	Normal Approximation of the Poisson	182
	Problems	182
Unit 7	Sampling Distributions and CLT	187
7.1	Sampling Distributions	188
7.1.1	Forming a Sampling Distribution of Means	188
7.1.2	Creating a Sampling Distribution of Sample Means	189
7.2	The Sampling Distribution of Sample Means	192
7.2.1	Central Limit Theorem	193
7.2.2	Constructing a Sampling Distribution of Sample Means	194
7.3	Application of the Sampling Distribution of Sample Means	199
7.3.1	Converting \bar{x} Information into z -scores	199
7.3.2	Distribution of \bar{x} and Increasing Individual Sample Size	200
7.4	Advanced Central Limit Theorem	202
7.4.1	Central Limit Theorem (Sample Mean)	203
7.4.2	Central Limit Theorem (Sample Sum)	203
	Problems	207

Part II Inferential Statistics

Unit 8 Introduction to Statistical Inferences	210
8.1 Point Estimation and Interval Estimation	211
8.1.1 Point Estimate	211
8.1.2 Interval Estimate	212
8.2 Estimation of Mean μ (σ Known)	214
8.2.1 The Principle of Constructing a Confidence Interval	214
8.2.2 Applications	216
8.2.3 Sample Size and Confidence Interval	217
8.3 Introduction to Hypothesis Testing	220
8.3.1 Null Hypothesis and Alternative Hypothesis	220
8.3.2 Four Possible Outcomes in a Hypothesis Test	222
8.4 Formulating the Statistical Null and Alternative Hypotheses	226
8.4.1 Writing Null and Alternative Hypothesis in One-Tailed Situation	226
8.4.2 Writing Null and Alternative Hypothesis in Two-Tailed Situation	227
8.5 Hypothesis Test of Mean μ (σ Known): A Probability-Value Approach	228
8.5.1 One-Tailed Hypothesis Test Using the p -Value Approach	229
8.5.2 Two-Tailed Hypothesis Test Using the p -Value Approach	233
8.5.3 Evaluating the p -Value Approach	234
8.6 Hypothesis Test of Mean μ (σ Known): A Classical Approach	235
8.6.1 One-Tailed Hypothesis Test Using the Classical Approach	236
8.6.2 Two-Tailed Hypothesis Test Using the Classical Approach	239
Problems	241
Unit 9 Inferences Involving One Population	246
9.1 Inferences about the Mean μ (σ Unknown)	247
9.1.1 Using the t-Distribution Table	249
9.1.2 Confidence Interval Procedure	251
9.1.3 Hypothesis-Testing Procedure	252
9.2 Inferences about the Binomial Probability of Success	258
9.2.1 Confidence Interval Procedure	259
9.2.2 Determining Sample Size	261
9.2.3 Hypothesis-Testing Procedure	263
9.3 Inferences about the Variance and Standard Deviation	268
9.3.1 Critical Values of Chi-Square	269
9.3.2 Hypothesis-Testing Procedure	270
Problems	279

Unit 10	Inferences Involving Two Populations	284
10.1	Dependent and Independent Samples	285
10.2	Inferences Concerning the Mean Difference Using Two Dependent Samples	287
10.2.1	Procedures and Assumptions for Inferences Involving Paired Data	287
10.2.2	Confidence Interval Procedure	288
10.2.3	Hypothesis-Testing Procedure	290
10.3	Inferences Concerning the Difference between Means Using Two Independent Samples	294
10.3.1	Confidence Interval Procedure	295
10.3.2	Hypothesis-Testing Procedure	297
10.4	Inferences Concerning the Difference between Proportions	301
10.4.1	Confidence Interval Procedure	303
10.4.2	Hypothesis-Testing Procedure	304
10.5	Inferences Concerning the Ratio of Variances Using Two Independent Samples	308
10.5.1	Writing for the Equality of Variances	308
10.5.2	Using the F -Distribution	309
10.5.3	One-Tailed Hypothesis Test for the Equality of Variances	310
10.5.4	Critical F -Values for One- and Two-Tailed Tests	313
	Problems	315
Unit 11	An Introduction to Simple Regression	321
11.1	Regression as a Best Fitting Line	322
11.1.1	Regression as a Best Fitting Line	322
11.1.2	Errors and Residuals	324
11.2	Interpreting OLS Estimates	326
11.3	Fitted Values and R^2 : Measuring the Fit of a Regression Model	328
11.4	Nonlinearity in Regression	331
	Reading English Materials	335
	Problems	336

Part III Statistical Methods and Data Science

Unit 12	Statistics and Data Science	339
12.1	Statistics and Data Science (I)	340
12.1.1	What is Data Science	340
12.1.2	Statistics and Data Science	340
12.2	Statistics and Data Science (II)	343
12.2.1	Statistics as Part of Data Science	343
12.2.2	The Modern Statistical Analysis Process	344

12.2.3	Statistician and Data Scientist	345
12.3	Statistical Thinking	348
12.3.1	What is Statistical Thinking	348
12.3.2	The Two Cultures of Statistical Modeling	348
12.3.3	A New Research Community	350
12.4	Distinguishing Analytics, Business Intelligence, Data Science	352
12.4.1	Analytics	352
12.4.2	Business Intelligence	355
12.4.3	Data Science	356
	Reading English Materials	359
	Problems	361
	Commonly Used Statistical Terms	362
	Appendix A Commonly Used Statistical Tables	367
	Appendix B Summary of Univariate Descriptive Statistics and Graphs for the Four Level of Measurement	379
	Appendix C Order of Magnitude of Data	380
	References	381

Part I Descriptive Statistics

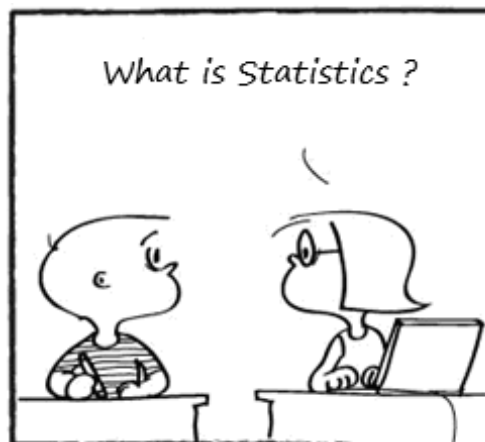
Most statistical work is concerned directly with the provision and implementation of methods for study design and for the analysis and interpretation of data.

—— D.R. Cox (He is one of the world's preeminent statisticians. His work on the proportional hazards regression model is one of the most-cited and most influential papers in modern statistics. In 2010 he won the Copley Medal of the Royal Society 'for his seminal contributions to the theory and application of statistics'. He is currently an Honorary Fellow at Nuffield College, Oxford.)



Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.

— H. G. Wells (1866—1946)



Unit 1

Statistics



1.1 What is Statistics?



1.2 The Language of Statistics



1.3 Measurability and Variability



1.4 Data Collection



1.5* Single-Stage Methods



1.6* Multistage Methods



1.7* Types of Statistical Study



1.8 The Process of a Statistical Study



Glossary



Reading English Materials



Problems

1.1 What is Statistics?

For a layman, ‘Statistics’ means numerical information expressed in quantitative terms. This information may relate to objects, subjects, activities, phenomena, or regions of space. As a matter of fact, data have no limits as to their reference, coverage, and scope. At the macro level, these are data on gross national product and shares of agriculture, manufacturing, and services in GDP (Gross Domestic Product). At the micro level, individual firms, howsoever small or large, produce extensive statistics on their operations. The annual reports of companies contain variety of data on sales, production, expenditure, inventories, capital employed, and other activities. These data are often field data, collected by employing scientific survey techniques. Unless regularly updated, such data are the product of a one-time effort and have limited use beyond the situation that may have called for their collection.

1.1.1 Meanings of Statistics

The word statistics has three different meanings (sense) which are discussed below: (1) Plural Sense; (2) Singular Sense; (3) Plural of the Word “Statistic”.

(1) **Plural Sense:** *In plural sense*, the word statistics refer to numerical facts and figures collected in a systematic manner with a definite purpose in any field of study. In this sense, statistics are also aggregates of facts which are expressed in numerical form. For example, statistics on industrial production, statistics on population growth of a country in different years etc.

(2) **Singular Sense:** *In singular sense*, it refers to the science comprising methods which are used in collection, analysis, interpretation and presentation of numerical data. These methods are used to draw conclusion about the population parameter.

For Example: If we want to have a study about the distribution of weights of students in a certain college. First of all, we will collect the information on the weights which may be obtained from the records of the college or we may collect from the students directly. The large number of weight figures will confuse the mind. In this situation we may arrange the weights in groups such as: “50 kg to 60 kg” “60 kg to 70 kg” and so on and find the number of students fall in each group. This step is called a presentation of data. We may still go further and compute the averages and some other measures which may give us complete description of the original data.

(3) **Plural of the Word “Statistic”:** The word statistics is used as the plural of the word “Statistic” which refers to a numerical quantity like mean, median, variance etc..., calculated from sample value.

For Example: If we select 15 students from a class of 80 students, measure their heights and find the average height. This average would be a statistic.

Definition 1 Two Means of Statistics

- **Statistics:** the *science* of collecting, describing, and interpreting data.
- **Statistics** are the *data* that describe or summarize something.

1.1.2 Definition of Statistics

Statistics like many other sciences is a developing discipline. It is not nothing static. It has gradually developed during last few centuries. In different times, it has been defined in different manners. Some definitions of the past look very strange today but those definitions had their place in their own time. Defining a subject has always been difficult task. A good definition of today may be discarded in future. It is difficult to define statistics. Some of the definitions are reproduced here:

The kings and rulers in the ancient times were interested in their manpower. They conducted census of population to get information about their population. They used information to calculate their strength and ability for wars. In those days statistics was defined as

“the science of kings, political and science of statecraft”.

A.L. Bowley has defined statistics as: (i) Statistics is the science of counting, (ii) Statistics may rightly be called the science of averages, and (iii) Statistics is the science of measurement of social organism regarded as a whole in all its manifestations.

A.L. Bowley defined statistics as *“statistics is the science of counting”*. This definition places the entries stress on counting only. A common man also thinks as if statistics is nothing but counting. This used to be the situation but very long time ago. Statistics today is not mere counting of people, counting of animals, counting of trees and counting of fighting force. It has now grown to a rich methods of data analysis and interpretation.

A.L. Bowley has also defined statistics as *“science of averages”*. This definition is very simple but it covers only some area of statistics. Average is very simple important in statistics. Experts are interested in average deaths rate, average birth rate, average increase in population, and average increase in per capita income, average increase in standard of living and cost of living, average development rate, average inflation rate, average production of rice per acre, average literacy rate and many other averages of different fields of practical life. But statistics is not limited to average only. There are many other statistical tools like measure of variation, measure of correlation, measures of independence etc... Thus this definition is weak and incomplete and has been buried in the past.

Professor Boddington has defined statistics as *“science of estimates and probabilities”*. This definition covers a major part of statistics. It is close to the modern statistics. But it is not complete because it stress only on probability. There are some areas of statistics in which probability is not used.

A definition due to W.I. King is “the science of statistics is the method of judging collection, natural or social phenomena from the results obtained from the analysis or enumeration or collection of estimates”. This definition is close to the modern statistics. But it does not cover the entire scope of modern statistics.

According to Professor Horace Secrist, statistics is the aggregate of facts, affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose, and placed in relation to each other.

According to Professor Ya Lu Chou, statistics is a method of decision making is the face of uncertainty on the basis of numerical data and calculated risks.

According to Professor *Wallis and Roberts*, statistics is not a body of substantive knowledge but a body of methods for obtaining knowledge.

Statistics both in the singular and the plural sense has been combined in the following definition which is accepted as the modern definition of statistics.

“Statistics are the numerical statement of facts capable of analysis and interpretation and the science of statistics is the study of the principles and the methods applied in collecting, presenting, analysis and interpreting the numerical data in any field of inquiry.”

From the above definitions, we can highlight the major characteristics of statistics as follows:

(i) *Statistics are the aggregates of facts.* It means a single figure is not statistics. For example, national income of a country for a single year is not statistics but the same for two or more years is statistics.

(ii) *Statistics are affected by a number of factors.* For example, sale of a product depends on a number of factors such as its price, quality, competition, the income of the consumers, and so on.

(iii) *Statistics must be reasonably accurate.* Wrong figures, if analyzed, will lead to erroneous conclusions. Hence, it is necessary that conclusions must be based on accurate figures.

(iv) *Statistics must be collected in a systematic manner.* If data are collected in a haphazard manner, they will not be reliable and will lead to misleading conclusions.

(v) Collected in a systematic manner for a pre-determined purpose.

(vi) Lastly, statistics should be placed in relation to each other. If one collects data unrelated to each other, then such data will be confusing and will not lead to any logical conclusions. Data should be comparable over time and over space.

1.1.3 Types of Statistics

The field of statistics can be roughly subdivided into two areas: descriptive statistics and inferential statistics. “**Descriptive statistics**” is what most people think of when they hear the word statistics. It includes the collection, presentation, and description of sample data. The term “**inferential statistics**” refers to the descriptive techniques and making decision and drawing conclusions about the population.

Statistics is more than just numbers: it is data, what is done to data, what is learned from the data, and the resulting conclusions. So, statistics is the science of collecting, describing, and interpreting data.

1.1.4 Applications of Statistics

Statisticians apply statistical thinking and methods to a wide variety of scientific, social, and business endeavors in such areas as astronomy, biology, education, economics, engineering, genetics, marketing, medicine, psychology, public health, sports, among many, see Figure 1.1. “The best thing about being a statistician is that you get to play in everyone else’s backyard.” (John Tukey, Bell Labs, Princeton University)

The uses of statistics are unlimited. It is much harder to name a field in which statistics is not used than it is to name one in which statistics plays an integral part. The following are a few

examples of how and where statistics are used:

- In education, descriptive statistics are frequently used to describe test results.
- In science, the data resulting from experiments must be collected and analyzed.
- In government, many kinds of statistical data are collected all the time. In fact, the U. S. government is probably the world's greatest collector of statistical data.
- In many other industries, such as in finance and economics and the like.

A very important part of the statistical process is that of studying the statistical results and formulating appropriate conclusion. These conclusions must then be communicated accurately—noting is gained from research unless the findings are shared with others. Statistics are being reported everywhere: newspapers, magazines, radio, and television. We read and hear about all kinds of new research results, especially in the health-related fields.

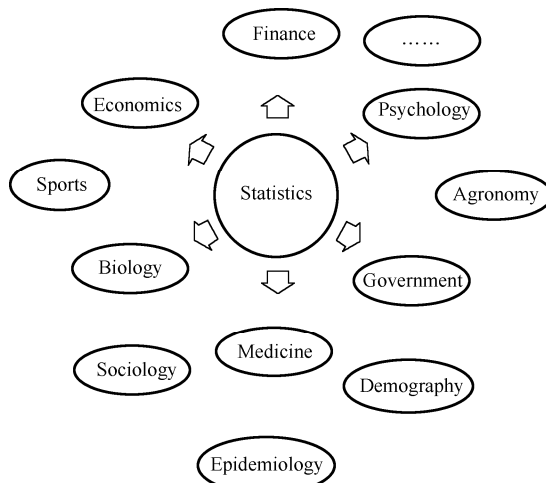


Figure 1.1 List of many fields of application of statistics

New Words and Expressions

statistics [stə'tɪstɪks] *n.* 统计, 统计学; 统计资料; 统计数字。“statistic”的复数, 这里 statistic 是统计上的一个术语, 意指统计量。

layman ['leɪmən] *n.* 门外汉, 外行; 俗人; 一般信徒

quantitative ['kwɒntɪtətɪv] *adj.* 定量的; 数量的, 数量上的

coverage ['kʌvərɪdʒ] *n.* 范围, 规模; 保险项目; (新闻) 报导

scope [skəʊp] *n.* (处理、研究事务的) 范围; 眼界, 见识; 广袤, 地域

howsoever [ˌhaʊsəʊ'evə] *adv.* 不管怎样, 无论如何

expenditure [ɪk'spendɪtʃə(r)] *n.* 花费, 支出; 费用, 经费

comprise [kəm'praɪz] *vt.* 包含, 包括; 由……组成; 由……构成

plural ['plʊərəl] *adj.* 复数的 *n.* 复数; 复数形式; 复数词

draw [drɔ:] *vt.* 画; 拉; 吸引 *vi.* 移动; 拔出剑; 汲取 *n.* 平局; 抽奖

draw conclusion 得出结论

ruler ['ru:lə(r)] *n.* 尺; 直尺; 统治者

manpower ['mænpaʊə(r)] *n.* 人力; 劳动力; 人力资源; 人手

manpower capital 人力资本; Manpower Inc 万宝盛华公司

statecraft ['stetkrɑ:ft] *n.* 管理国家的本领; 治国之道; 治国术

per capita [pə'kæpɪtə] *adj.* 每人; 按人分配的; 按人口平均

acre ['eɪkə(r)] *n.* 英亩; 土地, 耕地; [口]大量。per acre 每英亩

bury ['beri] *vt.* 埋葬; 隐藏, 埋藏, 遮盖, 掩蔽; 专心致志于, 埋头于, 沉溺于
inquiry [ɪn'kwaɪəri] *n.* 调查, 审查; 询问, 质问, 质询, 追究; 探究
pre-determined 预先安排, 预设; 预设的
erroneous [ɪ'rəʊniəs] *adj.* 错误的; 不正确的; 秕谬
haphazard [hæp'hæzəd] *adj.* 偶然的, 随意的; 无计划的; 任意的, 胡乱的
adv. 偶然地; 随意地; 无计划地; 杂乱无章地
endeavor [ɪn'devə] *vt. & vi.* 尝试, 试图; 尽力, 竭力 *n.* 努力, 尽力
astronomy [ə'strɒnəmi] *n.* 天文学
psychology [saɪ'kɒlədʒi] *n.* 心理学; 心理状态; 心理影响
medicine ['medsn] *n.* 医学; 药物; 有功效的东西
biology [baɪ'ɒlədʒi] *n.* 生物学; 生物。 *pl.* *biologies*
sociology [ˌsəʊsi'ɒlədʒi] *n.* 社会学; 群体生态学
demography [dɪ'mɒɡrəfi] *n.* 人口学; 人口统计; 人口统计学
sports [spɔ:t] *n.* 运动 (sport 的名词复数); 运动会; 突变; 娱乐
agronomy [ə'grɒnəmi] *n.* 农艺学, 农业; 作物学
epidemiology [ˌepɪ'dɪ:mi'ɒlədʒi] *n.* 流行病学; 传染病学
name [neɪm] *vt.* 确定; 决定; 给……取名。 *n.* 名字; 名声; 著名人物
penchant ['pɒʃ(ə)] *n.* (强烈的) 倾向, 爱好, 嗜好

Technical Terms

statistics *n.* 统计, 统计学; 统计资料; 统计数字
descriptive statistics 描述统计学
inferential statistics 推断统计学

Notes

1. 同义词辨析: compose, comprise, consist, constitute 这些动词均含“组成, 构成”之意。
compose: 正式用词, 多用被动态。指将两个或两个以上的人或物放到一起形成一个整体。
comprise: 正式用词, 指整体是由几个独立的部分所组成。
consist: consist 与 of 连用, 指一个整体由几个部分组成, 或由某些材料构成。
constitute: 正式用词, 指由某些部分组成一个整体或构成某物的基本成分。在句中, 主语表示事物的组成部分, 宾语表示事物的整体。

2. **Sir Arthur Lyon Bowley** (Bristol, 6 November 1869—Surrey, 21 January 1957) was an English statistician and economist who worked on economic statistics and pioneered the use of sampling techniques in social surveys.

3. **John Wilder Tukey** (June 16, 1915—July 26, 2000), an emeritus Princeton professor considered to be one of the most important contributors to modern statistics, and also was chemist, topologist, educator, consultant, information scientist, researcher, statistician, data analyst, executive.

Tukey developed many important tools of modern statistics and introduced concepts that were central to the creation of today's telecommunications technologies. In addition to his formidable research achievements, Tukey was known for his penchant for coining terms that reflected new ideas and techniques in the sciences and is credited with introducing the computer science terms "bit" (short for binary digit) and "software".

1.2 The language of Statistics

1.2.1 Population and Sample

To further continue our study, we need to "talk the talk". Statistics has its own jargon, terms beyond *descriptive statistics* and *inferential statistics*, that need to be defined and illustrated. The concept of a population is the most fundamental idea in statistics.

The **population** is the complete collection of individual or objects are of interest to the sample collector. The population of concern must be carefully defined and is considered fully defined only when its membership list of elements is specified. The set of "all students who have ever attended a U. S. college" is an example of a well-defined population.

Typically, we think of a population as a collection of people. However, in statistics the population could be a collection of animals, manufactured objects, whatever. For example, the set of all redwood trees in California could be a population.

There are two kinds of populations: finite and infinite. When the membership of population can be (or could be) physically listed, the population is said to be *finite*. When the membership is unlimited, the population is *infinite*. The books in your college library from a finite population; the OPAC (Online Public Access Catalog, the computerized card catalog) lists the exact membership. All the registered voters in the United States form a very large finite population; if necessary, a composite of all voter lists from all voting precincts across the United States could be compiled. On the other hand, the population of all people who might use aspirin and the population of all 40-watt light bulbs to be produced by Sylvania are infinite. Large populations are difficult to study; therefore, it is customary to select a sample, or a subset of a population, and study data in sample. A **sample** consists of the individuals, objects, or measurements selected from the population by the sample collector. See Figure 1.2.

Statisticians are interested in particular **variables** of a sample or a population. That is, they examine one or more characteristics of interest about each individual element of a population or sample. Things like age, hair color, height, and weight are variables. Each variable associated with one element of a population or sample has a value. That value, called the **data value**, may be a number, word, or symbol. For example, when Bill Jones entered college at age "23", his hair was "brown", he was "71 inches" tall, and he weighed "183 pounds". These four data values are the values for the variables as applied to Bill Jones.

The set of values collected from the variable from each of the elements that belong to the sample is called **data**. The set of 25 heights (or weight, ages, and hair colors) collected from students is an example of set of data. To collect a set of data, a statistician would do an **experiment**, which is a planned activity whose results yield a set of data. An experiment includes the activities for both selecting the elements and obtaining the data values.

The “average” age at time of admission for all students who have ever attended our college and the “proportion” of students who were older than 21 years of age when they entered college are examples of two population parameters. A **parameter** is a value that describes the entire population. Often a Greek letter is used to symbolize the name of a parameter. These symbols will be assigned as we study specific parameters.

For every parameter there is a corresponding sample statistic. The statistic is numerical value summarizing the sample data and describing the sample the same way the parameter describes the population.

The “average” height, found by using the set of 25 heights, is an example of sample statistic. A statistic is a value that describes a sample. Most sample statistics are found with the aid of formulas and are typically assigned symbolic names that are letters of the English alphabet (for example, \bar{x} , s , and r).

Definition 2

- The **population** in a statistical study is the *complete* set of people or things being studied.
- The **sample** is the subset of the population from which the raw data are actually obtained.

Definition 3

- **(Population) parameters** are specific characteristics of the population that a statistical study is designed to estimate.
- **(Sample) statistics** are numbers or observations that summarize the raw data.

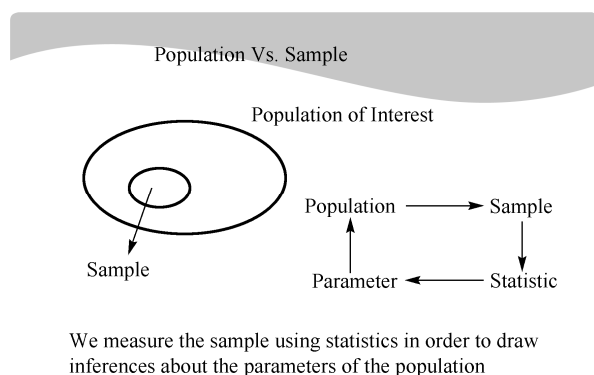


Figure 1.2 Population vs sample

FYI (for your information): Parameters describe the population; statistic describes the sample.

1.2.2 Kinds of Variables

There are basically two kinds of variables: (1) **qualitative variables** result in information that describes or categorizes an element of a population, and (2) **quantitative variables** result in information that quantifies an element of a population.

A sample of four hair-salon customers was surveyed for their “hair color”, “hometown”, and “level of satisfaction” with the results of their salon treatment. All three variables are examples of qualitative (attribute) variables because they describe some characteristic of the person, and all people with the same attribute belong to the same category. The data collected were {blonde, brown, black, brown}, {Brighton, Columbus, Albany, Jacksonville}, and {very satisfied, satisfied, somewhat satisfied}.

By contrast, the “total cost” of textbooks purchased by each student for this semester’s classes is an example of a quantitative (numerical) variable. A sample resulted in the following data: \$238.87, \$94.57, \$139.24. [To find the “average cost”, simply add the three numbers and divide by 3: $(238.87 + 94.57 + 139.24)/3 = \157.56 .] As you can see, arithmetic operations, such as addition and averaging, are meaningful for data that result from a quantitative variable (and would be useless in examining qualitative variables).

Each of these types of variables (qualitative and quantitative) can be further subdivided as illustrated in the following diagram, see Figure 1.3.

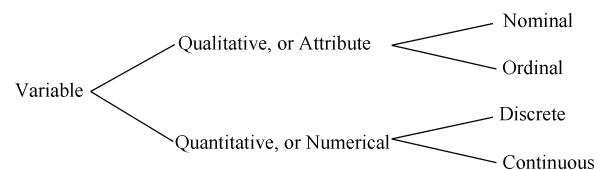


Figure 1.3 Kinds of variables

Qualitative variables may be characterized as nominal or ordinal. A **nominal variable** is a qualitative variable that characterizes (or describes, or names) an element of a population. Not only are arithmetic operations not meaningful for data that result from a nominal variable, but an order cannot be assigned to the categories.

In the survey of four hair-salon customers, two of the variables, “hair color” and “hometown”, are examples of nominal variables because both name some characteristic of the person, and it would be meaningless to find the sample average by adding and dividing by 4. For example, $(\text{blonde} + \text{brown} + \text{black} + \text{brown})/4$ is undefined. Furthermore, color of hair and hometown do not have an order to their categories.

An **ordinal variable** is a qualitative variable that incorporates an ordered position, or ranking. In the survey of four hair-salon customers, the variable “level of satisfaction” is an example of an ordinal variable because it does incorporate an ordered ranking: “Very satisfied” ranks ahead of “satisfied”, which ranks ahead of “somewhat satisfied”. Another illustration of an ordinal variable is the ranking of five landscape pictures according to someone’s preference: first choice, second choice, and so on.

Quantitative or numerical variables can also be subdivided into two classifications: *discrete variables* and *continuous variables*. A **discrete variable** is a quantitative variable that can assume a countable number of values. Intuitively, the discrete variable can assume any values corresponding to isolated points along a line interval. That is, there is a gap between any two values. By contrast,

a **continuous variable** is a quantitative variable that can assume an uncountable number of values. Intuitively, the continuous variable can assume any value along a line interval, including every possible value between any two values.

Definition 4

- **Discrete variable:** A quantitative variable that can assume a countable number of values.
- **Continuous variable:** A quantitative variable that can assume an uncountable number of values.

In many cases, the two types of variables can be distinguished by deciding whether the variables are related to a count or a measurement. The variable “number of courses for which you are currently registered” is an example of a discrete variable; the values of the variable may be found by counting the courses. (When we count, fractional values cannot occur; thus, there are gaps between the values that can occur.) The variable “weight of books and supplies you are carrying as you attend class today” is an example of a continuous random variable; the values of the variable may be found by measuring the weight. (When we measure, any fractional value can occur; thus, every value along the number line is possible.)

When trying to determine whether a variable is discrete or continuous, remember to look at the variable and think about the values that might occur. Do not look at only data values that have been recorded; they can be very misleading.

Consider the variable “judge’s score” at a figure skating competition. If we look at some scores that have previously occurred, 9.9, 7.4, 8.8, 10.0, and we see the presence of decimals, we might think that all fractions are possible and conclude that the variable is continuous. This is not true, however. A score of 9.134 is impossible; thus, there are gaps between the possible values and the variable is discrete.

Remember to inspect the individual variable and one individual data value, and you should have little trouble distinguishing among the various types of variables.

New Words and Expressions

jargon ['dʒɑ:gən] *n.* 行话; 行业术语; 黑话

population [ˌpɒpjə'leɪʃn] *n.* 人口; 全体居民; 特定[生物]种群

well-defined ['welɪ'daɪnd] *adj.* 定义明确的; 界限清楚的; 已知的

sample ['sɑ:mpl] *n.* 样品; 标本; (化验的) 取样; [信]信号瞬时值

vt. 取……的样品, 尤指用样品来检验; 品尝; 抽样调查 (通常用 sampling)

redwood ['redwud] *n.* (美国产的高大的) 红杉; 红杉木

finite ['famaɪt] *adj.* 有限的; [语]限定的; [数]有穷的, 有限的 *n.* 有限性; 有限的事物

infinite ['ɪnfɪnət] *adj.* 极大的; 无限的; 无穷尽的 *n.* 无限的事物; 无穷尽的事物; 上帝

precinct ['pri:sɪŋkt] *n.* [英] (教堂或大学) 的围地; [美]选区; 管辖区; 界限, 范围

compile *v.* 编译; 编写 (书、列表、报告等); 编纂

categorize ['kætəgəraɪz] *vt.* 把……归类, 把……分门别类

catalog ['kætəlɒg] *n.* 目录, 目录册; [美]大学概况一览; 登记, 记载; 产品样本
vt. 登记; 为……编目; 记载 *v.* 把……按目录分类; 把……编目

admission [əd'mɪʃn] *n.* 准许进入; 承认; 入场费

alphabet ['ælfəbet] *n.* 字母表; 字母系统; 入门, 初步

aspirin ['æsprɪn] *n.* 阿司匹林; 阿司匹林药片

variable ['veəriəbl] *n.* 可变因素, 变数, 变量
adj. 变化的, 可变的; [数]变量的; [生]变异的

salon ['sælɒn] *n.* 沙龙, 客厅; 画廊; 美术展览会。

hair-salon 美发沙龙, 美容院, 发廊

blonde [blɒnd] *adj.* (头发)亚麻色的, 金色的; 白皙的; 白肤金发碧眼的
n. 白肤金发碧眼女人

brown [braʊn] *adj.* 棕色的; 褐色的; 被晒黑的 *n.* 褐色; 棕色

isolated point *n.* 孤立, 孤立点; 孤点

misleading *adj.* 误导性的; 骗人的; 引入歧途的

figure skating 花样滑冰

decimal *adj.* 十进位的, 小数的 *n.* 小数

inspect [ɪn'spekt] *vt.* 检查, 检验; 视察, *vi.* 进行检查; 进行视察

Technical Terms

population 总体

sample 样本; 抽样

sampling 抽样; 取样

sample statistic 样本统计量

variable *n.* 变量

parameter 参数

Notes

1. OPAC (Online Public Access Catalog) 联机公共查询目录, 其含义是传统读者目录查询的自动化, 是一种通过网络查询馆藏信息资源的联机检索系统, 用户可以不受空间地点的限制, 查询各图书馆的 OPAC 资源。

2. 同义词辨析: alphabet, letter, character, script 均有“字母”之意。

alphabet 指整个字母系统或一种语言的字母表, 不表示单个字母。

letter 指单个的字母。

character 通常指汉语的方块字, 也指字符。

script 指书写或印刷的字母。

3. 常用的有关变量的词组

environment variable 环境变量; local variable 局部变量; random variable 随机变量;

qualitative variables 定性变量, 品质变量

quantitative variables 定量变量, 数量变量; 数量变数

nominal variable 名义变量, 名目变量, 列名型变量

ordinal variable 顺序变量, 次序变量, 有序变量

discrete variable 离散变量; continuous variable 连续变量

variable cost 可变成本

4. Figure Skate 花样的, 用于表现的, 包括常说的平地花式

Speed Skate or Racing Skate 速滑的

5. 有关小数的常用词组

decimal place 小数字; decimal fraction 纯小数

decimal point *n.* 小数点; decimal fraction *n.* 小数; decimal system *n.* 十进制

decimal digits *n.* 小数字数; decimal number (十进) 小数

infinite decimal 无穷小数; recurring decimal 循环小数

6. 同义词辨析 part, piece, section, division, portion, fraction, fragment, segment, share 均可表示“整体的一部分”之意。

part 含义非常广, 最普通用词, 指整体中可大可小的一部分, 也指整体中可分开的独立部分。

piece 指整体中的一些个体, 尤指从某个整体上分出来的一部分。

section 指整体中的分区, 部分与部分之间有显著界限。

division 通常指按类划分或分割而成的部分, 常含抽象意义。

portion 侧重从整体中所分配到的那一部分, 含一定的独立意义。

fraction 指包含在全体中的一部分, 暗指微不足道的一部分。

fragment 指因破裂、分割等产生的支离破碎、不规则的一部分。

segment 指某物的特定部分或自然形成的部分, 也指线形物品的一段。

share 指共有的东西中应占有的一部分。

1.3 Measurability and Variability

For example, we take a carton of a favorite candy bar and weigh each bar individually. We observe that each of the 24 candy bars weighs $7/8$ ounce, to the nearest $1/8$ ounce. Does this mean that the bars are all identical in weight? Not really! Suppose we were to weigh them on an analytical balance that weighs to the nearest ten-thousandth of an ounce. Now the 24 weights will most likely show **variability**.

It does not matter what the response variable is; there will most likely be variability in the data if the tool of measurement is precise enough. One of the primary objectives of statistical analysis is measuring variability. For example, in the study of quality control, measuring variability is absolutely essential. Controlling (or reducing) the variability in a manufacturing process is a field all its own namely, statistical process control.

Definition 5

■ **Variable (or response variable):** A characteristic of interest about each individual element of a population or sample.

Example 1.1 Applying the Basic Terms

A statistics student is interested in finding out something about the average dollar value of cars owned by the faculty members of our college. Eight of the terms just described can be identified in this situation.

1. The *population* is the collection of all cars owned by all faculty members at our college.
2. A *sample* is any subset of that population. For example, the cars owned by members of the mathematics department is a sample.
3. The *variable* is the “dollar value” of each individual car.
4. One *data value* is the dollar value of a particular car. Mr. Jones’s car, for example, is valued at \$9,400.
5. The *data* are the set of values that correspond to the sample obtained (9,400; 8,700; 15,950; …).
6. The *experiment* consists of the methods used to select the cars that form the sample and to determine the value of each car in the sample. It could be carried out by questioning each member of the mathematics department, or in other ways.
7. The *parameter* about which we are seeking information is the “average” value of all cars in the population.
8. The *statistic* that will be found is the “average” value of the cars in the sample.

Note: If a second sample were to be taken, it would result in a different set of people being selected—say, the English department—and therefore a different value would be anticipated for the statistic “average value”. The average value for “all faculty-owned cars” would not change, however.

FYI: Parameters are fixed in value, whereas statistics vary in value.

New Words and Expressions

candy ['kændi] *n.* 糖果; 冰糖; 巧克力 *adj.* 花哨的; 甜言蜜语的

ounce [aʊns] *n.* 盎司; 一点儿; 雪豹

variability [ˌveəriə'bɪləti] *n.* 变异性, 变化性, 易变, 变化的倾向; 变率

question *vt.* 问 (某人) 问题; 表示怀疑 *n.* 问题; 疑问; 议题

average ['ævərɪdʒ] *adj.* 平均的; 平常的; (价值、比率等的) 平均数的

n. 平均水平; (速度等的) 平均率; 平均估价

vt. [数学] 计算……的平均值; 分摊; 按比例 (或平均) 分配 (利润等)

anticipate [æn'tɪsɪpeɪt] *vt.* 预感; 预见; 先于……行动 *vi.* 过早地提出; 预言; 预测

Technical Terms

variability 变异性, 易变性

Notes

1. quality control *n.* 质量管理、质量控制。quality control engineer 质量管理工程师

2. statistical process control 统计过程控制, 简称 SPC, 是应用统计技术对过程中的各个阶段进行评估和监控, 建立并保持过程处于可接受的且稳定的水平, 从而保证产品与服务符合规定的要求的一种质量管理技术。SPC 源于 20 世纪 20 年代, 以美国 Shewhart 博士发明的控制图标志。自创立以来, 即在工业和服务等行业得到推广应用, 自 20 世纪 50 年代以来, SPC 在日本工业界的大量推广应用对日本产品质量的崛起起到了至关重要的作用; 20 世纪 80 年代以后, 世界许多大公司纷纷在自己内部积极推广应用 SPC, 对供应商也提出相应要求。

3. 词根词缀 后缀: -ability 表名词, “能……; 性质”

adaptability 适应性: adapt 适应+ability 能……; 性质→*n.* 适应性

dependability 可依赖性: depend 依靠, 依赖+ability 能……; 性质→*n.* 可依赖性

inflammability 易燃性: inflame 使燃烧+ability 能……; 性质→*n.* 易燃性

maintainability 可维护性: maintain 维修+ability 能……; 性质→*n.* 可维护性

portability 可携带, 轻便: port 拿, 运+ability 能……; 性质→*n.* 可携带、轻便

useability 可用性: use 用+ability 能……; 性质→*n.* 可用性

variability 可变性: vary 变化+ability 能……; 性质→*n.* 可变性

4. 同义词辨析: medium, median, average 这些形容词均含“中等的, 平均的, 适中的”之意。

medium 指按照某种标准来说是适中或中等的。这种标准可通过仪器测量而来, 也可能是凭经验而得出。

median 指中间位置的, 统计学上指处于中间位置的一个数。

average 通常用来形容优劣难分的平庸或折中情况, 也指理论上的平均标准。

5. 英文统计学书籍中, 经常出现“Average”和“Mean”, 翻译成中文都是“均值”或“平均数”的意思。其实, 这两个统计术语有不同的含义。

“Average”这个术语可以有三种表达方式, 分别是“Mean”、“Median”和“Mode”。具体来说, “Mean”的准确定义是指一组数的“算术平均值”或“算术平均数”; “Median”是指一组数的“中值”或“中位数”(即将这组数按从大到小或从小到大的顺序排列, 排在最中间的那个数即“中值”或“中位数”; 如果这组数有偶数个数, 则“中值”或“中位数”为按上述顺序排列的最中间的两个数的算术平均值); 而“Mode”则指一组数中的“众数”(即这组数中出现最多的那个数)。

在统计学中, 作为总体参数 (Population Parameter) 的“总体均值”只用“Mean”来表示 (Population Mean: μ); 而作为样本统计量 (Sample Statistic) 的“样本均值”, 则一般用“Average”来表示 (Sample Average: \bar{X}), 有时也可用“Mean”来表示 (Sample Mean: \bar{X})。

1.4 Data Collection

It is important to obtain “good data” because the inferences ultimately made will be based on the statistics obtained from these data. These inferences are only as good as the data.

Although it is relatively easy to define “good data” as data that accurately represent the population from which they were taken, it is not easy to guarantee that a particular sampling method will produce “good data”. As statisticians, we need to be on guard against **biased sampling methods** that produce data that systematically differ from the sampled population. We need to use sampling (data collection) methods that will produce data that are representative of the population and are *unbiased*, that is, not biased.

Definition 6

■ **Biased sampling method:** A sampling method that produces data that systematically differ from the sampled population.

Two commonly used sampling methods that often result in biased samples are the *convenience* and *volunteer samples*. A convenience sample, sometimes called a *grab* sample, occurs when items are chosen arbitrarily and in an unstructured manner from a population, whereas a volunteer sample consists of results collected from those elements of the population that chose to contribute the needed information on their own initiative.

Did you ever buy a basket of fruit at the market based on the “good appearance” of the fruit on top, only to later discover the rest of the fruit was not as fresh? It was too inconvenient to inspect the bottom fruit, so you trusted a convenience sample. Has your teacher used your class as a sample from which to gather data? As a group, the class was quite convenient, but is it truly representative of the school’s population? (Consider the differences among day, evening, and/or weekend students; type of course; etc.)

Have you ever mailed back your responses to a magazine survey? Under what conditions did (would) you take the time to complete such a questionnaire? Most people’s immediate attitude is to ignore the survey. Those with strong feelings will make the effort to respond; therefore, representative samples should not be expected when volunteer samples are collected.

1.4.1 The Data Collection Process

The collection of data for statistical analysis is an involved process and includes the following steps:

(I) Define the objectives of the survey or study. Examples: compare the effectiveness of a new drug to the effectiveness of the standard drug; estimate the average household income in the United States.

(II) Define the variable and the population of interest. Examples: length of recovery time for patients suffering from a particular disease; total income for households in the United States.

(III) Define the data collection and data measuring schemes. This includes sampling frame, sampling procedures, sample size, and the data measuring device (questionnaire, telephone, and so on).

(IV) Collect your sample. Select the subjects to be sampled and collect the data.

(V) Review the sampling process upon completion of collection. Often an analyst is stuck

with data already collected, possibly even data collected for other purposes, which makes it impossible to determine whether the data are “good”. Using approved techniques to collect your own data is much preferred. Although this text is concerned chiefly with various data analysis techniques, you should be aware of the concerns of data collection.

Two methods commonly used to collect data are *experiments* and *observational studies*. In an experiment, the investigator controls or modifies the environment and observes the effect on the variable under study. We often read about laboratory results obtained by using white rats to test different doses of a new medication and its effect on blood pressure. The experimental treatments were designed specifically to obtain the data needed to study the effect on the variable. In an observational study, the investigator does not modify the environment and does not control the process being observed. The data are obtained by sampling some of the population of interest. Surveys are observational studies of people.

If every element in the population can be listed, or enumerated, and observed, then a census is compiled. However, censuses are seldom used because they are often difficult and time-consuming to compile, and therefore very expensive. Imagine the task of compiling a census of every person who is a potential client at a brokerage firm. In situations similar to this, a *sample survey* is usually conducted.

1.4.2 Sampling Frame and Elements

When selecting a sample for a survey, it is necessary to construct a **sampling frame**, or a list, or set, of the elements belonging to the population from which the sample will be drawn. Ideally, the sampling frame should be identical to the population, with every element of the population included once and only once. In this case, a census would become the sampling frame. In other situations, a census may not be so easy to obtain, because a complete list is not available. Lists of registered voters or the telephone directory are sometimes used as sampling frames of the general public. Depending on the nature of the information being sought, the list of registered voters or the telephone directory may or may not serve as an unbiased sampling frame. Because only the elements in the frame have a chance to be selected as part of the sample, it is important that the sampling frame be representative of the population.

Once a representative sampling frame has been established, we proceed with selecting the sample elements from the sampling frame. This selection process is called the sample design. There are many different types of sample designs; however, they all fit into two categories: *judgment samples* and *probability samples*. **Judgment samples** are samples that are selected on the basis of being judged “typical”.

When a judgment sample is collected, the person selecting the sample chooses items that he or she thinks are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector’s judgment. This is not an acceptable statistical procedure.

Probability samples are samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of

the sample. The inferences that will be studied later in this textbook are based on the assumption that our sample data are obtained using a probability sample. There are many ways to design probability samples. We will look at two of them, single-stage methods and multistage methods, and learn about a few of the many specific designs that are possible. See Figure 1.4 and Figure 1.5.

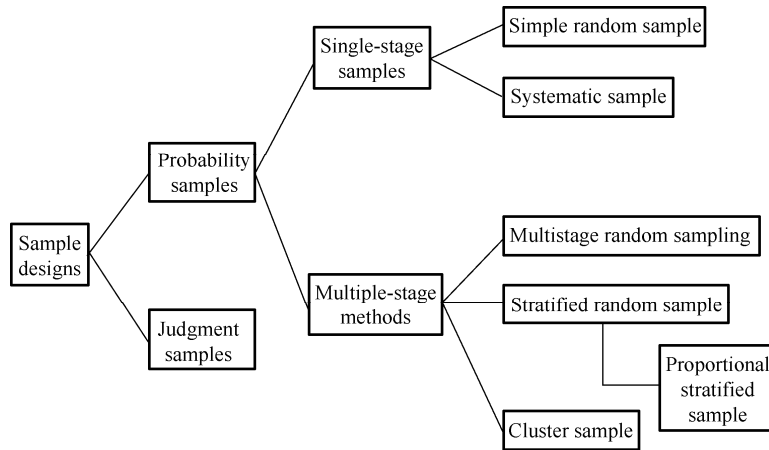


Figure 1.4 Sample designs

Definition 7

- **Judgment samples:** Samples that are selected on the basis of being judged “typical”.
- **Probability samples:** Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

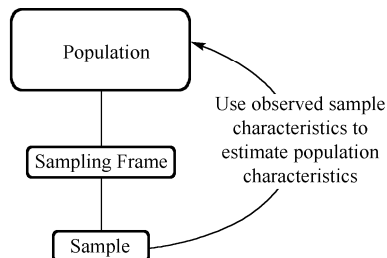
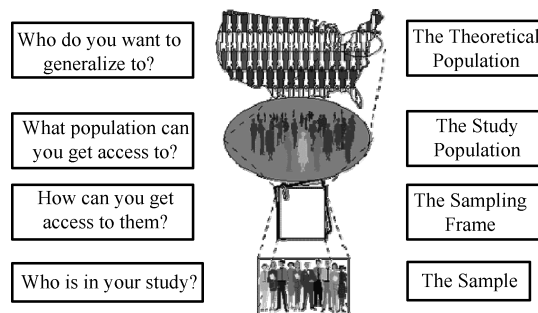


Figure 1.5 Sample designs and sample

New Words and Expressions

- ultimately ['ʌltɪmətli] *adv.* 最后，最终；基本上；根本
- representative [ˌreprɪˈzentətɪv] *n.* 代表；议员；类似物
adj. 典型的；有代表性的；代议制的
- volunteer [ˌvɒlənˈtɪə(r)] *n.* 志愿者，志愿兵；[军]义勇军；[植]自生植物
adj. 自愿的，志愿的
- grab [græb] *n.* 不法所得；被抓住的人；抓取装置 *vt. & vi.* 抢先，抢占
- initiative [ɪˈnɪʃətɪv] *n.* 主动性；主动精神；倡议；主动权 *adj.* 自发的；创始的；初步的
- arbitrarily [ˈɑːbɪtrəli] *adv.* 任意地；武断地；反复无常地；肆意地
- questionnaire [ˌkwɛstʃəˈneə(r)] *n.* 调查表；调查问卷
- ignore [ɪgˈnɔː(r)] *vt.* 忽视，不顾；[法律]驳回（诉讼）
- completion [kəmˈpliːʃn] *n.* 完成，结束；实现；[数]求全法；期满
- stick [stɪk] *vt.* 容忍；产生作用；（尤指迅速或随手）放置；阻延或推迟
n. 棍棒，棍枝；枝条；操纵杆；球棍
- stuck [stʌk] *adj.* 动不了的；被卡住的；被……缠住的；被……难住的
be stuck 卡住了；be stuck with 不得不；get stuck 遇到困难
- enumerate [ɪˈnjuːməreɪt] *vt.* 列举，枚举，数
- unbiased [ʌnˈbaɪəst] *adj.* 无偏见的，不偏不倚的，公正的；持平
- soundness [saʊndnəs] *n.* 完全坚固，公正，稳固

Technical Terms

- sample size 样本量，样本容量
- biased sampling method 有偏的抽样方法
- convenience samples 便利样本
- volunteer samples 自愿样本
- experiments study 实验研究
- observational study 观察研究，观测研究
- sampling frame 样本框
- judgment samples 判断抽样
- probability samples 概率抽样

Notes

1. Ideally, the sampling frame should be identical to the population, see Figure 1.6.
2. 同义词辨析：examine, inspect, investigate, scan 动词都有“调查，检查”之意。
examine 最普通用词，指粗略查看，也指仔细观察或调查以确定事物的性质、功能、特点等。

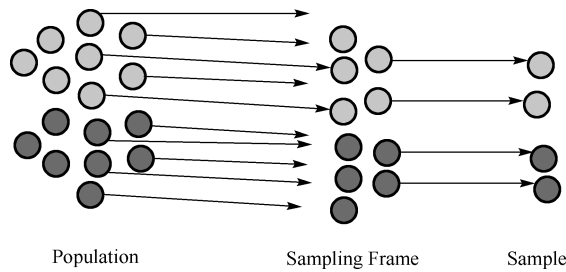


Figure 1.6 Sampling frame = population

inspect 侧重按一定质量标准检查某物，找出不足或不同之处。

investigate 指为发现事实真相或了解掌握情况而进行深入细致的现场考察。

scan 原意是仔细地检查分析，现用于指细看或浏览。

1.5* Single-Stage Methods

Single-stage sampling is a sample design in which the elements of the sampling frame are treated equally and there is no subdividing or partitioning of the frame. Two single-stage designs that statisticians use are the simple random sample and the systematic sample.

Definition 8

- **Single-stage sampling:** A sample design in which the elements of the sampling frame are treated equally and there is no subdividing or partitioning of the frame.

1.5.1 Simple Random Sample

One of the most common single-stage probability sampling methods used to collect data is the simple random sample, or a sample selected in such a way that every element in the population or sampling frame has an equal probability of being chosen. Equivalently, all samples of size n have an equal chance of being selected.

Definition 9

- **Simple random sample:** A sample selected in such a way that every element in the population or sampling frame has an equal probability of being chosen. Equivalently, all samples of size n have an equal chance of being selected.

Note: Random samples are obtained either by sampling with replacement from a finite population or by from an infinite population. See Figure 1.7.

Inherent in the concept of randomness is the idea that the next result (or occurrence) is not predictable. When a random sample is drawn, every effort must be made to ensure that each element has an equal probability of being selected and that the next result does not become predictable. The proper procedure for selecting a simple random sample requires the use of random numbers. Mistakes are commonly made because the term *random* (equal chance) is confused with haphazard (without pattern).

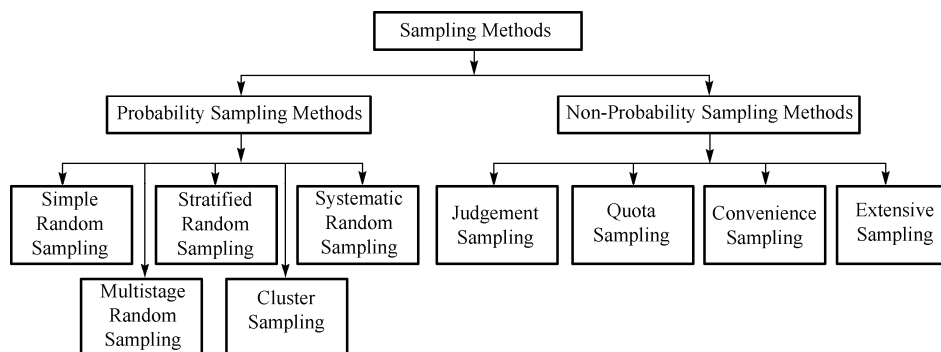


Figure 1.7 Sampling methods

To select a simple random sample, first assign an identifying number to each element in the sampling frame. This is usually done sequentially using the same number of digits for each element. Then using random numbers with the same number of digits, select as many numbers as are needed for the sample size desired. Each numbered element in the sampling frame that corresponds to a selected random number is chosen for the sample.

For example, imagine that the admissions office at your college wishes to estimate the current “average” cost of textbooks per semester, per student. The population of interest is the “currently enrolled student body”, and the variable is the “total amount spent for textbooks” by each student this semester. Because a random sample is desired, the dean of admissions has obtained a computer list of this semester’s full-time enrollment. Say there were 4,265 student names on the list and the dean numbered the students 0001, 0002, 0003, and so on, up to 4265; then, using four-digit random numbers, the dean identified a sample; 1288, 2177, 1952, 2463, 1644, 1004, and so on, were selected.

Why create a random sample? Because a simple random sample is the first step toward an unbiased sample. Random samples are required for most of the statistical procedures presented in this book. Without a random design, the conclusions we draw from the statistical procedures may not be reliable.

1.5.2 Systematic Sample

In concept, the simple random sample is the simplest of the probability sampling techniques, but it is seldom used in practice because it often is an inefficient technique. One of the easiest methods to use for approximating a simple random sample is the **systematic sampling method**, which involves selecting every k th item of the sampling frame, starting from a first element, which is randomly selected from the first k elements.

To select an x percent (%) systematic sample, we will need to randomly select 1 element from every $\frac{100}{x}$ elements. After the first element is randomly located within the first $\frac{100}{x}$ elements, we proceed to select every $\frac{100}{x}$ th item thereafter until we have the desired number of data values for our sample.

For example, if we desire a 3% systematic sample, we would locate the first item by randomly selecting an integer between 1 and 33 ($\frac{100}{x} = \frac{100}{3} = 33.33$, which when rounded becomes 33).

Suppose 23 be randomly selected. This means that our first data value is obtained from the subject in the 23rd position in the sampling frame. The second data value will come from the subject in the 56th ($23 + 33 - 56$) position; the third, from the 89th ($56 + 33$); and so on, until our sample is complete.

Definition 10

■ **Systematic sample:** A sample in which every k th item of the sampling frame is selected, starting from a first element, which is randomly selected from the first k elements.

The systematic technique is easy to describe and execute; however, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature. For example, a systematic sample of every k th house along a long street might result in a sample disproportional with regard to houses on corner lots. The resulting information would likely be biased if the purpose for sampling is to learn about support for a proposed sidewalk tax. In these situations the results may not approximate a simple random sample.

New Words and Expressions

random ['rændəm] *adj.* 任意的; 随机的; 胡乱的 *n.* 随意; 偶然的行动

randomly ['rændəmlɪ] *adv.* 随机地, 随便地, 未加计划地

randomness ['rændəmnəs] *n.* 随意, 无安排; 随机性

haphazard [hæp'hæzəd] *adj.* 偶然的, 随意的; 无计划的; 任意的
n. 偶然事件; 偶然性; 任意性

inherent [ɪn'hɪərənt] *adj.* 固有的, 内在的; 天生

digit ['dɪdʒɪt] *n.* 数字; 手指, 足趾; 一指宽

dean [di:n] *n.* (大学的) 学院院长, 系主任; 教务长; 学监

randomization [,rændəmaɪ'zeɪʃən] *n.* 随机化, 随机选择

inefficient [ɪnɪ'fɪʃnt] *adj.* 无效率的, 无能的; 不称职的; 徒劳的

approximate [ə'prɒksɪmət] *vi.* 接近于; 近似于; 逼近于 *vt.* 靠近; 使接近; 使结合

round [raʊnd] *vt. & vi.* 使成圆形; 绕行; 拐过, 绕过; 把……四舍五入

n. 圆, 圆形; 循环; 圆形物, 球状物。词组 round off 表示四舍五入

integer ['ɪntɪdʒə(r)] *n.* 整数

rounded ['raʊndɪd] *adj.* 圆形的; 丰满的; 全面的; 四舍五入的

cyclical *adj.* 循环的; 周期的; 环状的。

cyclic economy 循环经济; cyclic transformation 循环变换

disproportional [dɪsprə'pɔ:ʃənəl] *adj.* 不相称的, 不成比例的

corner *n.* 角; 拐角; 街角; 墙角 *v.* 转弯; 使(人或动物)走投无路; 逼……入绝境

English corner 英语角; corner solution 角点解

sidewalk ['saɪdwɔ:k] *n.* 人行道, 小路, 行人路。sidewalk artist 路边肖像画家

Technical Terms

single-stage sampling 单阶段抽样

simple random sample 简单随机抽样

systematic sample 系统抽样, 等距抽样

sampling without replacement 无放回抽样、不放回抽样

sampling with replacement 有放回抽样、放回抽样、重复抽样

Notes

1. 常用词组: at random 随机的; random number 随机数; random number generator 随机数发生器; random number table 随机数表、随机数目录表。

固定词组: stochastic process 随机过程; stochastic model 随机模型; stochastic noise 随机噪声。通常, random 更口语化, 使用频率要高一些, 简单地说, random 更偏向于生活中事物的无规则, 比如抽奖和随手拿书; stochastic 是一个书面语, 多用在科技领域。

2. 同义词辨析: ① inherent, essential 形容词均有“内在的, 本质的”之意。

inherent: 指物体本身固有的、不能与该物体分割的某种特性。

essential: 指决定所属事物存在的关键因素。

② angle, corner 两个名词都有“角”之意。

angle 几何学术语, 指两条直线相交而成的角, 也可引申指看问题的方面或角度。

corner 多指物体的棱角或房间、街道的角落。

3. admission office 招生办; full-time enrollment 全职注册生, 全职注册生人数。

part-time enrollment 兼职注册生, 兼职注册生人数。

4. 数学上常用数字表示法:

null, zero, nought, nil 零; operator 运算符; digit 数字; number 数; natural number 自然数; positive integer 正整数; integer 整数; negative integer 负整数; decimal point 小数点; decimal or decimal number 小数; fraction 分数

5. zero, naught, nil, null 的区别 (数字 0 在英国英语中有几个不同的名称, 在美国英语中通常用 zero 表示):

① zero ['zɪərəʊ] *n.* (数字) 0; 零度。 *adj.* 全无的, 没有的

作为数字, 用于精确的科学、医学和经济方面, 亦用以表示温度, absolute zero 绝对零度; zero hour 零点时刻。比如: zero inflation/ growth/ profit。

② naught [nɔ:t] *n.* 零; 乌有; 泡影

在英国英语中用以表示数字、年龄等, 比如: A million is written with six noughts. 再比如:

The car goes from nought to sixty in ten seconds.

③ nil [nɪl] *n.* 无, 零; 零分

在某些谈及体育的语境, 用以表示团队比赛 (如足球赛) 的比分等, 比如: England beat

Poland two-nil at Wembley. 再比如：The final score was one nil.(1-0)。

④ null [nʌl] *adj.* 零值的，等于零的

比如：a null result/output 毫无结果；零输出。再比如：The contract was declared as null and void. 合同被宣布无效。null and void 是惯用语，通常用以表示选举、协议等无法律效力的。
null hypothesis (统计学) 零假设、原假设。

1.6* Multistage Methods

When sampling very large populations, sometimes it is necessary to use a *multistage sampling* design to approximate random sampling. **Multistage random sampling** is a sample design in which the elements of the sampling frame are subdivided and the sample is chosen in more than one stage.

Multistage sampling designs often start by dividing a very large population into subpopulations on the basis of some characteristic. These subpopulations are called *strata*. These smaller, easier-to-work-with strata can then be sampled separately. One such sample design is the **stratified random sampling method**. This method produces a sample by stratifying the population, or sampling frame, and then selecting a number of items from each of the strata by means of a simple random sampling technique.

Definition 11

- **Multistage random sampling:** A sample design in which the elements of the sampling frame are subdivided and the sample is chosen in more than one stage,
- **Stratified random sample:** A sample obtained by stratifying the population, or sampling frame, and then selecting a number of items from each of the strata by means of a simple random sampling technique.

A stratified random sample results when the population, or sampling frame, is subdivided into various strata, usually some already occurring natural subdivision, and then a subsample is drawn from each of these strata. These subsamples may be drawn from the various strata by using random or systematic methods. The subsamples are summarized separately first and then combined to draw conclusions about the entire population.

When a population with several strata is sampled, we often require that the number of items collected from each stratum be proportional to the size of the strata; this method is called a proportional stratified sampling. After stratifying the population or sampling frame, the researcher then selects a number of items in proportion to the size of the strata from each strata by means of a simple random sampling technique.

Definition 12

- **Proportional stratified sample:** A sample obtained by stratifying the population, or sampling frame, and then selecting a number of items in proportion to the size of the strata from each strata by means of a simple random sampling technique.

A convenient way to express the idea of proportional sampling is to establish a quota. For example, the quota, “1 for every 150” directs you to select 1 data value for each 150 elements in each strata. That way, the size of the strata determines the size of the subsample from that strata. The subsamples are summarized separately and then combined to draw conclusions about the entire population.

A cluster sample is another multistage design. A **cluster sample** is obtained by stratifying the population, or sampling frame, and then selecting some or all of the items from some, but not all, of the strata. The cluster sample uses either random or systematic methods to select the strata (clusters) to be sampled (first stage) and then uses either random or systematic methods to select elements from each identified cluster (second stage). The cluster sampling method also allows the possibility of selecting all of the elements from each identified cluster. Either way, the subsamples are summarized separately, and the information then combined. See Figure 1.8.

Definition 13

- **Cluster sample:** A sample obtained by stratifying the population, or sampling frame, and then selecting some or all of the items from some, but not all, of the strata.

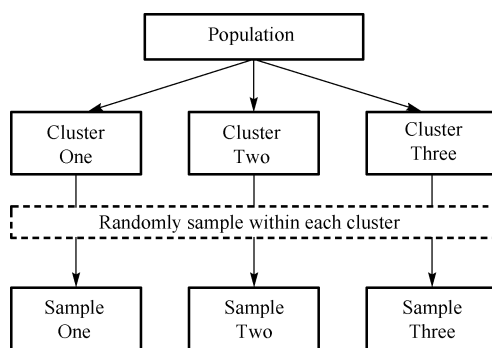


Figure 1.8 Cluster sampling

To illustrate a possible multistage random sampling process, consider that a sample is needed from a large country. In the first stage, the country is divided into smaller regions, such as states, and a random sample of these states is selected. In the second stage, a random sample of smaller areas within the selected states (counties) is then chosen. In the third stage, a random sample of even smaller areas (townships) is taken within each county. Finally in the fourth stage, if these townships are sufficiently small for the purposes of the study, the researcher might continue by collecting simple random samples from each of the identified townships. This would mean the entire sample was made up of several “local” subsamples identified as a result of the several stages.

Sample design is not a simple matter; many colleges and universities offer separate courses in sample surveying and experimental design. The topic of survey sampling is a complete textbook in itself. It is intended that the preceding information will provide you with an overview of sampling and put its role in perspective.

New Words and Expressions

- strata ['strɑ:tə] *n.* 层; 岩层(stratum 的名词复数); 地层; 社会阶层
township ['taʊnʃɪp] *n.* 小镇; 镇区; 种族
ethnic township 民族乡; township enterprises 乡镇企业; township head 乡长
quota ['kwəʊtə] *n.* (正式限定的) 定量, 定额; 配额; 指标
subsample ['sʌbsɑ:mpl] *n.* 子样本; (从样品中再抽取) 子样品, 二次抽样样品
vt. 对……作二次抽样
perspective [pə'spektɪv] *n.* 透镜, 望远镜; 观点, 看法; 远景; 洞察力
adj. (按照) 透视画法的; 透视的

Technical Terms

- multistage sampling 对阶段抽样
stratified random sample 分层随机抽样
proportional stratified sample 比例分层抽样
cluster sample 整群抽样

Notes

配额抽样 quota sampling 也称“定额抽样”, 是指调查人员将调查总体样本按一定标志分类或分层, 确定各类(层)单位的样本数额, 在配额内任意抽选样本的抽样方式。

配额抽样和分层随机抽样既有相似之处, 也有很大区别。配额抽样和分层随机抽样有相似的地方, 都是事先对总体中所有单位按其属性、特征分类, 这些属性、特征称为“控制特性”。例如, 市场调查消费者的性别、年龄、收入、职业、文化程度等。然后, 按各个控制特性, 分配样本数额。它与分层抽样又有区别, 分层抽样是按随机原则在层内抽选样本, 而配额抽样则是由调查人员在配额内主观判断选定样本。

1.7* Types of Statistical Study

Broadly speaking, most statistical studies fall into one of two categories: observational studies and experiments, see Figure 1.9. Nielsen's studies of television viewing are observational because they are designed to observe the television-viewing behavior of the people in its 5000 sample homes. Note that observational studies may still involve some interaction. For example, an opinion poll is observational, even though researchers may conduct in-depth interviews, because the poll's goal is to learn (observe) people's opinions, not to change them. Similarly, a study in which individuals in the sample are weighed is also observational, because the measurement process records (observes) but does not change a person's weight.

In contrast, consider a medical study designed to test whether large doses of vitamin C can help prevent colds. To conduct this study, the researchers must ask some people in the sample to take large doses of vitamin C. This type of statistical study is called **an experiment**, because some participants receive a treatment (in this case, vitamin C) that they would not otherwise receive.

Definition 14 Two Basic Types of Statistical Study

- In an **observational study**, researchers observe or measure characteristics of the sample members but do not attempt to influence or modify these characteristics.
- In an **experiment study**, researchers apply a treatment to some or all of the sample members and then look to see whether the treatment has any effects.

It is difficult to determine whether an experimental treatment works unless you compare groups that receive the treatment to groups that don't. In the vitamin C study, for example, researchers might create two groups of people: a treatment group that takes large doses of vitamin C and a control group that does not take vitamin C. The researchers can then look for differences in the numbers of colds among people in the two groups. Having a control group is usually crucial to interpreting the results of experiments.

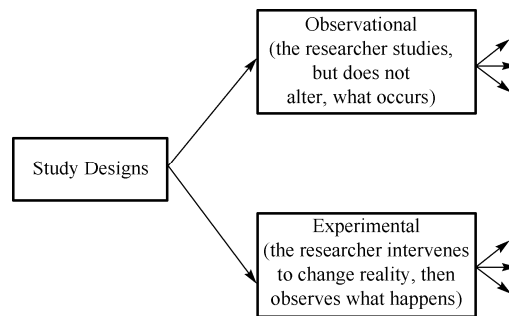


Figure 1.9 Two types of statistical study

In an experiment, it is very important for the treatment and control groups to be alike in all respects except for the treatment. For example, if the treatment group consisted of active people with good diets and the control group consisted of sedentary people with poor diets, we could not attribute any differences in colds to vitamin C alone. To avoid this type of problem, assignments to the control and treatment groups must be done randomly.

Definition 15 Treatment and Control Group

- The **treatment group** in an experiment is the group of sample members who receive the treatment being tested.
- The **control group** in an experiment is the group of sample members who do not receive the treatment being tested.

It is important for the treatment and control groups to be selected randomly and to be alike in all respects except for the treatment.

The Placebo Effect and Blinding

For experiments involving people, using a treatment and a control group might not be enough to get reliable results. The problem is that people can be affected by their beliefs as well as by real treatments. For example, stress and other psychological factors have been shown to affect resistance to colds. If people taking vitamin C get fewer colds than people who don't, we can't conclude that the vitamin C was responsible. It might be that people stayed healthier because they believed that vitamin C works. Therefore, people in the control group should be given a **placebo**—in this case, pills that look like vitamin C pills but don't actually contain vitamin C. As long as the participants don't know whether they are in the treatment or control group (that is, whether they got the real pills or the placebo), any effect arising from psychological factors—known as a **placebo effect**—should affect both groups equally. Then, if people in the vitamin C group get fewer colds than people in the control group, we have evidence that vitamin C really works.

Definition 16 Placebo and Placebo Effect

- A **placebo** lacks the active ingredients of a treatment being tested in a study, but is identical in appearance to the treatment. Thus, study participants cannot distinguish the placebo from the real treatment.
- The **placebo effect** refers to the situation in which patients improve simply because they believe they are receiving a useful treatment.

In statistical terminology, the practice of keeping people in the dark about who is in the treatment group and who is in the control group is called blinding. A single-blind experiment is one in which the participants don't know which group they belong to, but the experimenters (the people administering the treatment) do know.

Using a placebo is one way to create a single-blind experiment. Sometimes, a single-blind experiment can still be unreliable if the experimenters can subtly influence outcomes. For example, in an experiment that involves interviews, the experimenters might speak differently to people who received the real treatment than to those who received the placebo. This type of problem can be avoided by making the experiment double-blind, which means neither the participants nor the experimenters know who belongs to each group, see Figure 1.10. (Of course, someone must keep track of the two groups in order to evaluate the results at the end. In typical double-blind experiments, researchers hire experimenters to make any necessary contact with the participants.)

Definition 17 Single-blind and Double-blind

- An **experiment is single-blind** if the participants do not know whether they are members of the treatment group or members of the control group, but the experimenters do know.
- An **experiment is double-blind** if neither the participants nor the experimenters (people administering the treatment) know who belongs to the treatment group and who belongs to the control group.

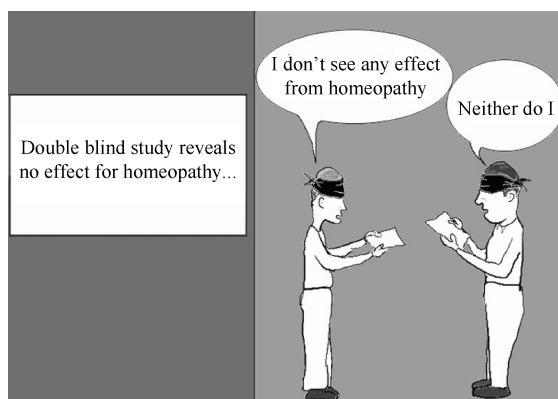


Figure 1.10 Double-blind study

Experiment or Observational Study

SURGICAL INFECTION IS A MATTER OF TIME

Many surgical patients fail to get timely doses of the right medications, raising the risk of infection, researchers write in the *Archives of Surgery*. Of 30 million operations performed each year in the USA, about 2% are complicated by an onside infection, the report says. The study of 34,000 surgical patients at nearly 3,000 hospitals in 2001 found that only 56% got prophylactic medications within an hour of surgery, when they can be effective.

(Source: USA Today, February 22, 2005)

This study is an example of an observational study. The researchers did not modify or try to control the environment. They observed what was happening and wrote up their findings.

New Words and Expressions

interview ['ɪntəvju:] *n.* 采访; 面试; 接见; 会谈 *vt.* 采访; 访问; 会见; (私下) 提问

in-depth interview 深度访谈, 深入访谈法, 深度访谈形式

vitamin ['vɪtəmin] *n.* 维生素; 维他命

participant [pɑ:'tɪsɪpənt] *n.* 参加者, 参与者; 与会代表; 关系者 *adj.* 参加的; 有关系的

sedentary ['sedntri] *adj.* 坐着的; (指工作等) 坐着干的; 案头的; (指人) 不爱活动的

sedentary farming 定居耕种; sedentary soil 原地土壤

diet ['daɪət] *n.* 日常饮食; 规定饮食 *vi.* 节食; 进规定饮食

attribute [ə'trɪbjʊ:t] *vt.* 认为……是; 把……归于; 认为某事属于某人

cold [kəʊld] *adj.* 寒冷的; 冷淡的, 无情的; 失去知觉的 *n.* 寒冷; 感冒, 伤风

alike [ə'lɪk] *adj.* 同样的, 相似的 *adv.* 同样地; 两者都; 类似于

psychological [ˌsaɪkə'lɒdʒɪkəl] *adj.* 心理的, 精神的

psychological health 心理健康

placebo [plə'si:bəʊ] *n.* 安慰剂; 安慰剂效应

dark [dɑ:k] *adj.* 黑暗的; 忧郁的; 神秘的 *n.* 黑暗; 暗色; 暗处

blind [blaɪnd] *adj.* 失明的；盲目的；供盲人用的
vt. 弄瞎，使失明；蒙蔽，欺瞒；使变暗
n. 窗帘，百叶窗；用以蒙蔽人的言行；借口；

administer *vt.* 管理；治理（国家）；给予；执行
vi. 执行遗产管理人的职责；给予帮助；担当管理人

unreliable [ˌʌnrɪˈlaɪəbl] *adj.* 不可靠的，靠不住的；不能信任的

evaluate [ɪˈvæljuet] *vt.* 评价；求……的值或数；对……评价；[数学、逻辑学]求……的数值
vi. 评价，估价

medication [ˌmedɪˈkeɪʃn] *n.* 药物，药剂；药物治疗；加入药物

infection [ɪnˈfekʃn] *n.* [医]传染，感染；传染病，染毒物；影响

onside [ˌɒnˈsaɪd] *adj.* （在某些运动比赛中）没有越位的；支持的；赞同的

prophylactic [ˌprɒfɪˈlæktɪk] *adj.* 预防（性）的 *n.* 预防剂；预防用品；预防法；避孕用品

surgery ['sɜ:dʒəri] *n.* 外科学，外科手术；手术室；诊所；诊断时间

modify ['mɒdɪfaɪ] *vi.* 被修饰；修改 *vt.* 改变；减轻，减缓；[语]修饰

homeopathy [ˌhəʊmiˈɒpəθi] [医]顺势医疗论，类似医疗论

Technical Terms

observational study 观察研究，观测研究
treatment group 治疗组
control group 对照组，控制组
placebo effect 安慰剂效应
single-blinding experiment 单盲实验
double-blinding experiment 双盲实验

Notes

1. 同义词辨析：estimate, appraise, assess, evaluate, value, rate 这些动词均有“估价，估计”之意。
estimate 通常指由个人作出的主观估价。
appraise 指以专家身份作了最终精确的估价。
assess 原意指为确定交多少税而估计，引申指通过估价以便更好地利用。
evaluate 与 appraise 相似，指使判断绝对准确，但多用于对人物的某方面的评价，很少用于评价某物的市场价值。
value 侧重指一般人对某物的价值或价格所作的估计。
rate 专指评定价值等级的高低。
2. write up 全部写出。The show received a good write-up. 演出获得了好评。

1.8 The Process of a Statistical Study

Statistical studies are conducted in many different ways and for many different purposes, but they all share a few characteristics. To get the basic ideas, consider the *Nielsen ratings*（尼尔森收

收视率，又称尼尔森收视率统计），which are used to estimate the numbers of people watching various television shows. These ratings are used, for example, to determine the most popular television show of the week.

Nielsen's goal is to draw conclusions about the viewing habits of all Americans. In the language of statistics, we say that Nielsen is interested in the *population* of all Americans. The characteristics of this population that Nielsen seeks to learn—such as the number of people watching each television show—are called *population parameters*.

Nielsen seeks to learn about the population of all Americans by studying a much smaller *sample* of Americans in depth. More specifically, Nielsen has devices (called “people meters”) attached to televisions in 5000 homes, so the people who live in these homes make up the sample of Americans that Nielsen studies, see Figure 1.11. The individual measurements that Nielsen collects from the sample, such as who is watching each show at each time, constitute the *raw data*. Nielsen then consolidates these raw data into a set of numbers that characterize the sample, such as the percentage of young male viewers watching *Lost*. These numbers are called sample statistics.

Suppose the *Nielsen ratings* tell you that *Lost* was last week's most popular show, with 22 million viewers, see Figure 1.12. You probably know that no one actually counted all 22 million people. But you may be surprised to learn that the Nielsen ratings are based on the television-viewing habits of people in only 5000 homes. To understand how Nielsen can draw a conclusion about millions of Americans from 5000 homes, we need to investigate the principles behind statistical research.

Because Nielsen does not study the entire population of all Americans, it cannot actually measure any population parameters. Instead, the company tries to infer reasonable values for population parameters from the sample statistics (which it did measure).



Figure 1.11 People meters

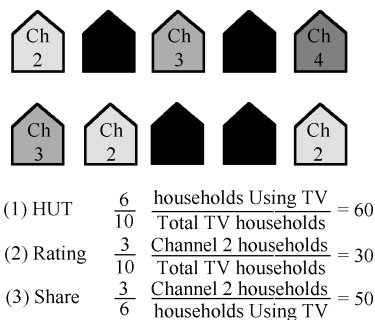


Figure 1.12 Nielsen ratings

The process of inference is simple in principle, though it must be carried out with great care. For example, suppose Nielsen finds that 7% of the people in its sample watched *Lost*. If this sample accurately represents the entire population of all Americans, then Nielsen can infer that approximately 7% of all Americans watched the show. In other words, the sample statistic of 7% is used as an estimate for the population parameter. (By using statistical techniques that we'll discuss in Unit 6, Nielsen can also estimate the uncertainty in the inferred population parameters.)

Once Nielsen has estimates of the population parameters, it can draw general conclusions about what Americans were watching. The process used by Nielsen Media Research is similar to that used in many statistical studies. Figure 1.13 summarizes the general relationships among a population, a sample, the sample statistics, and the population parameters.

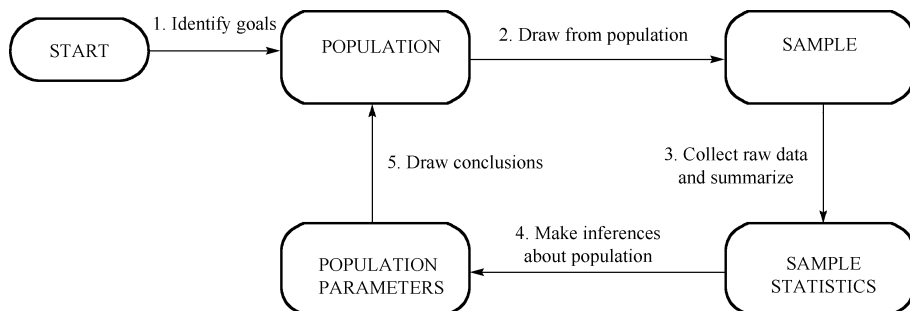


Figure 1.13 Elements of a statistical study

◇ Basic Steps in a Statistical Study ◇

1. State the goal of your study precisely. That is, determine the population you want to study and exactly what you'd like to learn about it.
2. Choose a representative sample from the population.
3. Collect raw data from the sample and summarize these data by finding sample statistics of interest.
4. Use the sample statistics to infer the population parameters.
5. Draw conclusions: Determine what you learned and whether you achieved your goal.

New Words and Expressions

raw data [rɔ:'deɪtə] 原始[未处理]数据, 原始材料, 素材

consolidate [kən'sɒlɪdeɪt] vt. 把……合成一体; 巩固, 加强; 统一 vi. 统一; 合并; 联合

Lost 迷失是一部美国电视连续剧, 讲述从澳大利亚悉尼飞往美国洛杉矶的海洋航空公司 815 航班在南太平洋一个神秘热带小岛上坠毁后生还者的生活和经历的事。

Technical Terms

Nielsen Media Research 尼尔森媒体研究; 尼尔森媒体研究公司

Nielsen ratings 尼尔森收视率, 又称尼尔森收视率统计

people meter 直译尼尔森的人员测量仪; 收视纪录器; 收视仪

Notes

1. learn from sb./sth. 学; 学习; 学到; 学会

How much did you learn from the course.

2. gain from 从某物中受益；从……得到；获益良多

The answer depends on what you think you gain from innovation and lose from crises.

3. share with 合伙，分享，承担，跟……看法一致，与……分享，共用

Let me share with you one of the funny story.

Glossary

Population: A collection, or set, of individuals, objects, or events whose properties are to be analyzed.

Finite Population: A population whose membership can or could be physically listed.

Infinite Population: A population whose membership is unlimited.

Sample: A subset of a population.

Variable (or response variable): A characteristic of interest about each individual element of a population or sample.

Data value: The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

Data: The set of values collected from the variable from each of the elements that belong to the sample.

Experiment: A planned activity whose results yield a set of data.

Parameter: A numerical value summarizing all the data of an entire population.

Statistic: A numerical value summarizing the sample data.

Qualitative (or attribute or categorical) Variable: A variable that describes or categorizes an element of a population.

Quantitative (or numerical) Variable: A variable that quantifies an element of a population.

Nominal Variable: A qualitative variable that characterizes (or describes, or names) an element of a population.

Ordinal Variable: A qualitative variable that incorporates an ordered position, or ranking.

Passage 1. What is Statistics?

When many people hear the word “statistics”, they think of either sports-related numbers or the college class they took and barely passed. While statistics can be thought about in these terms, there is more to the relationship between you and statistics than you probably imagine.

So, what is statistics? Several informal definitions are offered in the book *A Career in Statistics: Beyond the Numbers* by Gerald Hahn and Necip Doganaksoy:

- The science of learning from (or making sense out of) data—J. Kettenring
- The theory and methods of extracting information from observational data for solving real-world problems—C. R. Rao
- The science of uncertainty—D. J. Hand
- The quintessential interdisciplinary science—S. McNulty
- The art of telling a story with [numerical] data—L. Gaines

Statistics is used around the world by governments, political parties, civil servants, financial companies, opinion-polling firms, social researchers, news organizations, and so much more.

Statisticians, the scientists who collect and analyze data, work in many areas that touch your life, including the following:

○ Medicine ○ Economics ○ Agriculture ○ Business ○ Law enforcement ○ Weather forecasting ...

Statistics is becoming more critical as academia, businesses, and governments come to rely on data-driven decisions, greatly expanding the demand for statisticians. Statistics has become the universal language of the sciences.

Passage 2. From Data to Foresight

Data: Raw, unprocessed facts and /or figures, often obtained via use of measurement instruments.

Information: Data that has been processed and structured, adding context and increased meaning.

Knowledge: Ability to use information tactically and strategically to achieve specified objectives.

Wisdom: Ability to use select objectives that are consistent with and supportive of a general set of values, such as human values.

Foresight: Ability to accurately predict outcomes of one’s proposed decisions and actions.

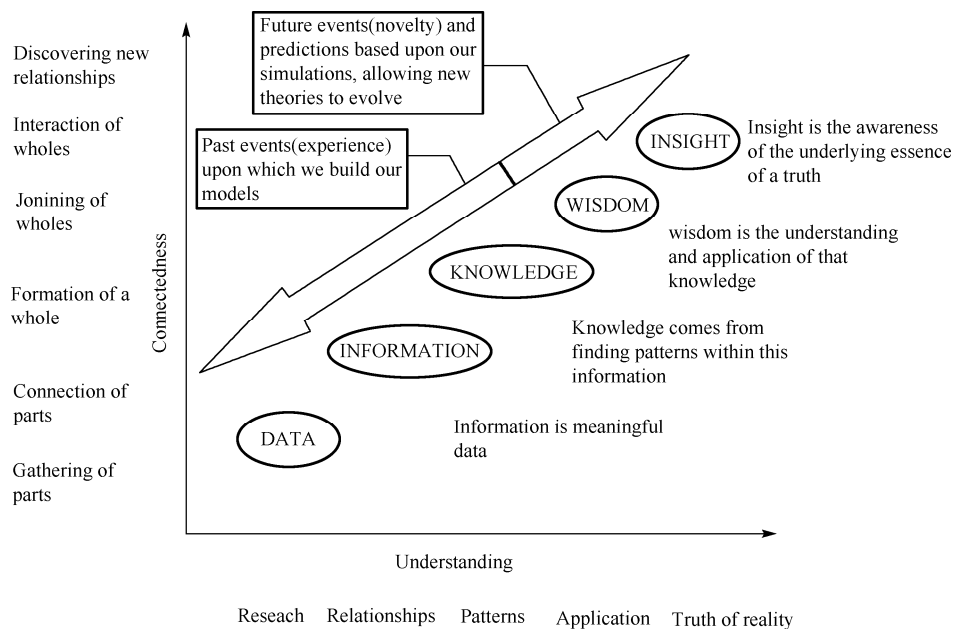


Figure 1.14 From data to foresight

Problems

1.1 Please give out some definitions of statistics (minimum by 4 authors) as explained by various distinguished authors.

1.2 International Communications Research (ICR, 国际传播研究所) conducted the 2004 National Spring Cleaning Survey for The Soap and Detergent Association. ICR questioned 3000 American male and female heads of household regarding their house cleaning attitudes, The survey has a margin of error of plus or minus 5%.

- What is the population?
- How many people were polled?
- What information was obtained from each person?

THOSE HARD TO CLEAN PLACES

Cleaning windows is rated the most difficult household task by more than one-third of adults.

Behind the TV	24%
Don't know	8%
Tops of shelves	16%
Under couch	12%
Venetian blinds	35%
Wooden floors	5%

Data from Anne R. Carey and Gia Kereselidze, *USA Today*: Source: Swifter

- Using the information given, estimate the number of surveyed adults who think cleaning under the couch is the most difficult cleaning job,
- What do you think the “margin of error of plus or minus 5%” means?

f. How would you use the “margin of error” in estimating the percentage of all adults who think that Venetian blinds are the hardest to clean?

(注释: margin of error 误差范围。Venetian blinds 威尼斯式软百叶帘。)

1.3 Identify each of the following as examples of (1) nominal, (2) ordinal, (3) discrete, or (4) continuous variables:

- a. A poll of registered voters as to which candidate they support
- b. The length of time required for a wound to heal when a new medicine is being used
- c. The number of televisions within a household
- d. The distance first-year college women can kick a football
- e. The number of pages per job coming off a computer printer
- f. The kind of tree used as a Christmas tree

1.4 Select 10 students currently enrolled at your college and collect data for these three variables:

X: number of courses enrolled in

Y: total cost of textbooks and supplies for courses

Z: method of payment used for textbooks and supplies

- a. What is the population?
- b. Is the population finite or infinite?
- c. What is the sample?
- d. Classify the three variables as nominal, ordinal, discrete, or continuous.

1.5 A coin-operated coffee vending machine dispenses, on the average, 6 oz of coffee per cup. Can this statement be true of a vending machine that occasionally dispenses only enough to fill the cup half full (say, 4 oz)? Explain.

(注释: vending machine 投币式自动售货机。oz 是符号 ounce 的缩写, 中文称为“盎司”是英制计量单位, 一种质量单位[ounce (缩写 oz)], 英语 ounce 的音译。英制重量计量单位, 为一磅的十六分之一, 等于 28.3495 克。)

1.6 A wholesale food distributor in a large metropolitan area would like to test the demand for a new food product. He distributes food through five large supermarket chains. The food distributor selects a sample of stores located in areas where he believes the shoppers are receptive to trying new products. What type of sampling does this represent?

1.7 a. What body of the federal government illustrates a stratified sampling of the people? (A random selection process is not used.)

b. What body of the federal government illustrates a proportional sampling of the people? (A random selection process is not used.)

1.8 Consider a simple population consisting of only the numbers 1, 2, and 3 (an unlimited number of each). Nine different samples of size 2 could be drawn from this population: (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3).

a. If the population consists of the numbers 1, 2, 3, and 4, list all the samples of size 2 that could possibly be selected.

b. If the population consists of the numbers 1, 2, and 3, list all the samples of size 3 that could possibly be selected.

1.9 Why is the random sample so important in statistics?

1.10 Describe in detail how you would select a 4% systematic sample of the adults in a nearby large city in order to complete a survey about a political issue.

1.11 What is meant by the saying “Garbage-in, garbage-out!” and how have computers increased the probability that studies may be victimized (victimize *vt.* 使受骗) by the adage (*n.* 谚语, 格言)?

Garbage-In, Garbage-out!

There is a great temptation to use the computer or calculator to analyze any and all sets of data and then to treat the results as though the statistics are correct. Remember the adage: “Garbage in equals garbage out.” This phrase became popular during the era of computer software’s explosive growth to reflect the fact that the solution or output of a software program can only be as good as the information being input to the software.

Responsible use of statistical methodology is very important. The burden is on the user to ensure that the appropriate methods are correctly applied and that accurate conclusions are drawn and communicated to others.



A picture is worth a thousand words.

— Frederick R. Barnard

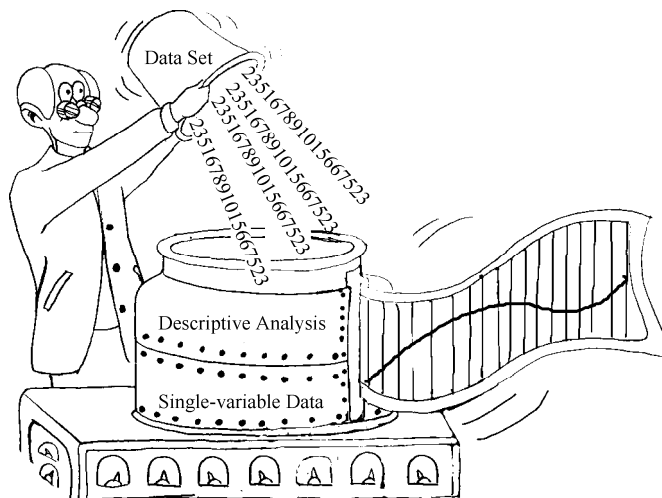
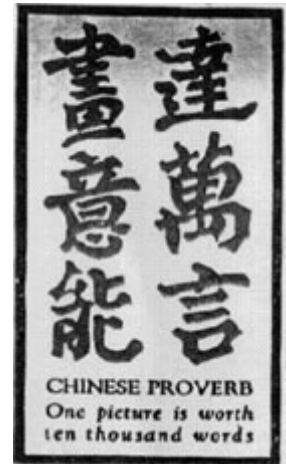
Origin

This phrase emerged in the USA in the early part of the 20th century. Its introduction is widely attributed to Frederick R. Barnard, who published a piece commending the effectiveness of graphics in advertising with the title “One look is worth a thousand words”, in *Printer’s Ink*, December 1921. Barnard claimed the phrase’s source to be oriental by adding “so said a famous Japanese philosopher, and he was right”.

Printer’s Ink printed another form of the phrase in March 1927, this time suggesting a Chinese origin:

“Chinese proverb. One picture is worth ten thousand words.”

The arbitrary escalation from ‘one thousand’ to ‘ten thousand’ and the switching from Japan to China as the source leads us to smell a rat with this derivation. In fact, Barnard didn’t introduce the phrase—his only contribution was the incorrect suggestion that the country of origin was Japan or China. This has led to another popular belief about the phrase, that is, that it was coined by Confucius. It might fit the Chinese-sounding ‘Confucius he say’ style, but the Chinese derivation was pure invention.



Unit 2

Descriptive Analysis of Single-Variable Data



2.1 Graphs, Pareto Diagrams and Stem-and-Leaf Displays



2.2 Frequency Distributions and Histograms



2.3 Measures of Central Tendency



2.4 Measures of Dispersion



2.5 Measures of Position



2.6 Interpreting and Understanding Standard Deviation



Glossary



Problems

2.1 Graphs, Pareto Diagrams, and Stem-and-Leaf Displays

Once the sample data have been collected, we must “get acquainted” with data.

One of the most helpful ways to become acquainted with the data is to use an initial exploratory data analysis technique that will result in a pictorial representation of the data. The display will visually reveal patterns of behavior of the variable being studied. There are several graphic (pictorial) ways to describe data. The type of data and the idea to be presented determine which method is used.

There is no single correct answer when constructing a graphic display. The analyst’s judgment and the circumstances surrounding the problem play a major role in the development of the graph.

2.1.1 Qualitative Data

Graphs can be used to summarize qualitative, or attribute, or categorical, data. **Circle graphs (pie diagrams)** show the amount of data that belong to each category as a proportional part of a circle. **Bar graphs** show the amount of data that belong to each category as a proportionally sized rectangular area. Any graphic representation used, regardless of type, needs to be completely self-explanatory. That includes a descriptive, meaningful title and proper identification of the quantities and variables involved. To appreciate the differences between these two types of graphical representations, let’s compare them by using the same data set to create one of each.

To get a better sense of what’s involved in graphing qualitative data, let’s consider an example about surgeries at a hospital. Table 2.1 lists the number of cases of each type of operation performed at General Hospital last year.

The data in Table 2.1 are displayed on a circle graph in Figure 2.1, with each type of operation represented by a relative proportion of the circle, found by dividing the number of cases by the total sample size namely, 498. The proportions are then reported as percentages (for example, 25% is 1/4 of the circle). Figure 2.2 displays the same “type of operation” data but in the form of a bar graph. Bar graphs of attribute data should be drawn with a space between bars of equal width.

Table 2.1 Operations Performed at General Hospital Last Year

Type of Operation	Number of Cases
Thoracic	20
Bones and joints	45
Eye, ear, nose, and throat	58
General	98
Abdominal	115
Urologic	74
Proctologic	65
Neurosurgery	23

Definition 1

- **Circle graphs (pie diagrams):** Graphs that show the amount of data belonging to each category as a proportional part of a circle.

Definition 2

- **Bar graphs:** Graphs that show the amount of data belonging to each category as a proportionally sized rectangular area.

When the bar graph is presented in the form of a Pareto diagram, it presents additional and very helpful information. That's because in a Pareto diagram the bars are arranged from the most numerous category to the least numerous category. A Pareto diagram also includes a line graph displaying the cumulative percentages and counts for the bars. The Pareto diagram is popular in quality-control applications. A Pareto diagram of types of defects will show the ones that have the greatest effect on the defective rate in order of effect. It is then easy to see which defects should be targeted in order to most effectively lower the defective rate.

Pareto diagrams can also be useful in evaluating crime statistics. The FBI reported the number of hate crimes by category for 1993 (*USA Today*, June 29, 1994). The Pareto diagram in Figure 2.3 shows the 6,746 categorized hate crimes, their percentages, and cumulative percentages.

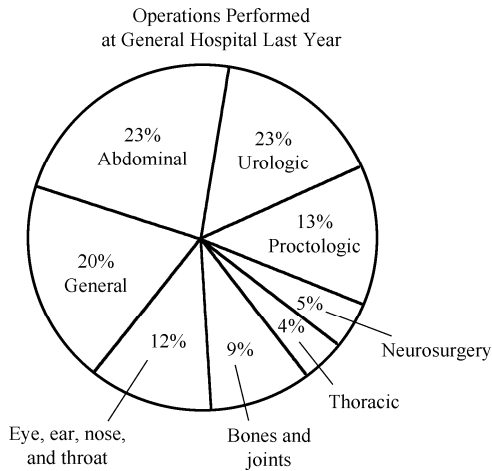


Figure 2.1 Circle graph

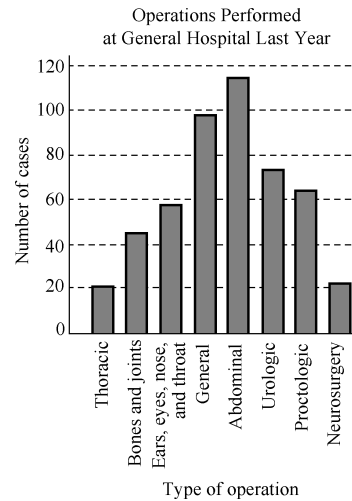


Figure 2.2 Bar graph

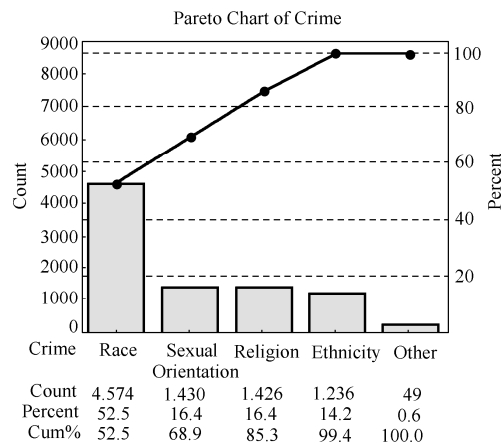


Figure 2.3 Pareto diagram

Definition 3

- **Pareto diagram:** A bar graph with the bars arranged from the most numerous category to the least numerous category. It includes a line graph displaying the cumulative percentages and counts for the bars.

2.1.2 Quantitative Data

One major reason for constructing a graph of quantitative data is to display its **distribution**, or the pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable. Two popular methods for displaying distribution of qualitative data are the dotplot and the stem-and-leaf display.

Definition 4

- **Distribution:** The pattern of variability displayed by the data of a variable. The distribution displays the frequency of each value of the variable.

Dotplot

One of the simplest graphs used to display a distribution is the dotplot. The dotplot displays the data of a sample by representing each data value with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale. The dotplot display is a convenient technique to use as you first begin to analyze the data. It results in a picture of the data as well as sorts the data into numerical order. (To sort data is to list the data in rank order according to numerical value.) Table 2.2 provides a sample of 19 exam grades randomly selected from a large class. Notice how the data in Figure 2.4 are “bunched” near the center and more “spread out” near the extremes.

Table 2.2 Sample of 19 Exam Grades

76	74	82	96	66	76	78	72	52	68
86	84	62	76	78	92	82	74	88	

Definition 5

- **Dotplot display:** Displays the data of a sample by representing each data with a dot positioned along a scale. This scale can be either horizontal or vertical. The frequency of the values is represented along the other scale.

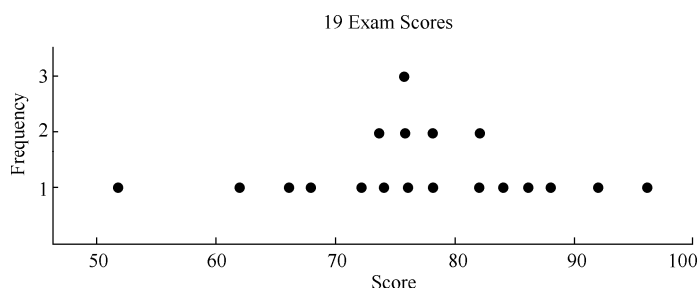


Figure 2.4 Dotplot

Stem-and-Leaf Display

In recent years a technique known as the stem-and-leaf display has become very popular for summarizing numerical data. It is a combination of a graphic technique and a sorting technique. These displays are simple to create and use, and they are well suited to computer applications. A stem-and-leaf display displays the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The leading digit(s) becomes the stem, and the trailing digit(s) becomes the leaf. The stems are located along the main axis, and a leaf for each data value is located so as to display the distribution of the data.

Definition 6

■ **Stem-and-leaf display:** A display of the data of a sample using the actual digits that make up the data values. Each numerical value is divided into two parts: The leading digit(s) becomes the stem, and the trailing digit(s) becomes the leaf. The stems are located along the main axis, and a leaf for each data value is located so as to display the distribution of the data.

Let's construct a stem-and-leaf display for the 19 exam scores from Table 2.2. At a quick glance we see that there are scores in the 50s, 60s, 70s, 80s, and 90s. Let's use the first digit of each score as the stem and the second digit as the leaf. Typically, the display is constructed vertically. We draw a vertical line and place the stems, in order, to the left of the line.

$$\begin{array}{c|c} 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array}$$

Next we place each leaf on its stem. This is done by placing the trailing digit on the right side of the vertical line opposite its corresponding leading digit. Our first data value is 76; 7 is the stem and 6 is the leaf. Thus, we place a 6 opposite the stem 7:

$$7 \mid 6$$

The next data value is 74, so a leaf of 4 is placed on the 7 stem next to the 6.

$$7 \mid 6 \quad 4$$

The next data value is 82, so a leaf of 2 is placed on the 8 stem.

$$\begin{array}{c|c} 7 & 6 \quad 4 \\ 8 & 2 \end{array}$$

We continue until each of the other 16 leaves is placed on the display. Figure 2.5A shows the resulting stem-and-leaf display; Figure 2.5B shows the completed stem-and-leaf display after the leaves have been ordered.

From Figure 2.5B, we see that the grades are centered around the 70s. In this case, all scores with the same tens digit were placed on the same branch, but this may not always be desired.

Suppose we reconstruct the display; this time instead of grouping ten possible values on each stem, let's group the values so that only five possible values could fall on each stem. Do you notice a difference in the appearance of Figure 2.6? The general shape is approximately symmetrical about the high 70s. Our information is a little more refined, but basically we see the same distribution.

19 Exam Scores	
5	2
6	6 8 2
7	6 4 6 8 2 6 8 4
8	2 6 4 2 8
9	6 2

Figure 2.5A Unfinished stem-and-leaf display

19 Exam Scores	
5	2
6	2 6 8
7	2 4 4 6 6 6 8 8
8	2 2 4 6 8
9	2 6

Figure 2.5B Final stem-and-leaf display

It is fairly typical of many variables to display a distribution that is concentrated (mounded) about a central value and then in some manner dispersed in one or both directions. Often a graphic display reveals something that the analyst may or may not have anticipated. The example that follows demonstrates what generally occurs when two populations are sampled together.

19 Exam Scores	
(50-54) 5	2
(55-59) 5	
(60-64) 6	2
(65-69) 6	6 8
(70-74) 7	2 4 4
(75-79) 7	6 6 6 8 8
(80-84) 8	2 2 4
(85-89) 8	6 8
(90-94) 9	2
(95-99) 9	6

Figure 2.6 Stem-and-leaf display

New Words and Expressions

pictorial [pɪk'tɔ:riəl] *adj.* 绘画的; 有图片的; 图画似的 *n.* 画报; 画页; 图画邮票

proportional [prə'pɔ:ʃənl] *adj.* 比例的, 成比例的; 相称的, 均衡的

n. [数]比例项, 比例量

proportionally [prə'pɔ:ʃənlɪ] *adv.* 按比例地, 相配合地, 适当地

proportionality constant 比例常数; proportionality coefficient 比例系数

diagram ['daɪəgræm] *n.* 图表; 图解; 示意图; [数]线图 *vt.* 用图表示; 图解

rectangular [rek'tæŋgjələ(r)] *adj.* [数]矩形的; 成直角的; 直角坐标的

rectangular distribution 矩阵分布

self-explanatory [self ɪk'splænətɹɪ] *adj.* 不解自明的, 明显的; 自解释
 operation [ˌɒpə'reɪʃn] *n.* 操作, 经营; 手术; [数]运算; 作用
 percentage [pə'sentɪdʒ] *n.* 百分比, 百分率; 比例, 部分; [数]百分法
 thoracic [θɔ:'ræsɪk] *adj.* 胸的 Thoracic Surgery 胸外科
 abdominal [æb'dɒmɪnl] *adj.* 腹部的 *n.* 腹肌 (常用作复数)
 urologic [juərəu'lɒdʒɪk] *adj.* 泌尿科学的, 泌尿道的 urological surgeon 泌尿外科医师
 proctology [prɒk'tɒlədʒɪ] *n.* 直肠病学
 proctologic [prɒk'tɒlədʒɪk] 直肠的; 直肠病学的
 neurosurgery ['njuərəʊsɜ:dʒəri] *n.* 神经外科 neurosurgery department 神经外科
 space [speɪs] *n.* 空间, 太空; 空白, 间隔; 空隙; 空格
 vt. 把……分隔开, 留间隔于……之间
 crime [kraɪm] *n.* 罪行, 犯罪; 罪恶 *v.* 指控犯罪; 判定犯罪; 处罚军事犯
 axis ['æksɪs] *n.* 轴, 轴线; 轴心国

Technical Terms

exploratory data analysis 探索性数据分析
 circle graph 扇形图, 圆形图, 圆图
 bar graph 条形图; 柱状图
 Pareto diagram 帕累托图, 排列图
 quality-control 质量管理
 stem-and-leaf display 茎叶图
 cumulative percentages 累积频率, 累计频率
 dotplot (有时也写作 dot plot) 点图; 二维点图; 点阵图
 trailing digits 尾部数字

Notes

1. 同义词辨析: drawing, illustration, cartoon, diagram, picture, sketch, painting, portrait 这些名词都表示“画, 图”之意。

drawing: 指用线条或色彩绘成的图画。

illustration: 指插入书页之间帮助说明的任何插图或图解。

cartoon: 指幽默或讽刺性漫画。

diagram: 多指科技书籍或文献中具有概括解说作用的图表、图样或略图。

picture: 指广义的“图画”, 现多用来指相片、画像。

sketch: 通常指只画出物体主要特征的图画。

painting: 指着色的画。

portrait: 指肖像, 只用于指人。

2. 同义词辨析: draft, outline, diagram, plot, sketch, blueprint 这些词既可作动词也可作名词用, 作动词时均有“绘制”之意; 作名词时都含“草图”之意。

draft: 用作动词时指按准确比例设计或打样; 作名词时专指精确的草图或草案。

outline: 主要给出事物要点或轮廓, 强调简化了的整体。

diagram: 侧重指用图形、图表等来说明。

plot: 可与 draft 和 diagram 换用, 但侧重于表示具体的点、面、部分或目标, 从而使相互关系以及和整体的关系得以明确。

sketch: 指用图、模型或语言描述来表示某一事物的整体情况。

blueprint: 主要指绘制蓝图、制定纲领或规划。这个词引申用来指详细而具体的行动计划。

3. 同义词辨析: percent, percentage 这两个名词均可表示“百分比”之意。

percent: 系拉丁语 per centum 的缩略, 通常和一个具体数字连用, 指具体的百分比。

percentage: 不受数字修饰, 不指具体的百分比, 通常用在一些形容词或起形容词作用的词之后, 也可单独使用。

2.2 Frequency Distributions and Histograms

2.2.1 Frequency Distribution

Sometimes we want to condense the data into a more manageable form. This can be accomplished by creating a **frequency distribution** that pairs the values of a variable with their frequency. Frequency distributions are often expressed in chart form.

To demonstrate the concept of a frequency distribution, let's use this set of data:

3 2 2 3 2 4 4 1 2 2
4 3 2 0 2 2 1 3 3 1

If we let x represent the variable, then we can use a frequency distribution to represent this set of data by listing the x values with their frequencies. For example, the value 1 occurs in the sample three times; therefore, the frequency for $x = 1$ is 3. The complete set of data is shown in the frequency distribution in Table 2.4.

Definition 7

■ **Frequency distribution:** A listing, often expressed in chart form, that pairs values of a variable with their frequency.

The **frequency** f is the number of times the value x occurs in the sample. Table 2.3 is an *ungrouped frequency distribution*—“ungrouped” because each value of x in the distribution stands alone. When a large set of data has many different x values instead of a few repeated values, as in the previous example, we can group the values into a set of classes and construct a *grouped frequency distribution*. The stem-and-leaf display in Figure 2.5B shows, in picture form, a grouped frequency distribution. Each stem represents a class. The number of leaves on each stem is the

same as the frequency for that same class (sometimes called a bin). The data represented in Figure 2.5B are listed as a grouped frequency distribution in Table 2.4.

Definition 8

■ **Frequency:** The number of times the value x occurs in the sample.

Table 2.3 Ungrouped Frequency Distribution

x	f
0	1
1	3
2	8
3	5
4	3

Table 2.4 Grouped Frequency Distribution

		Class	Frequency
50 or more to less than 60	→	$50 \leq x < 60$	1
60 or more to less than 70	→	$60 \leq x < 70$	3
70 or more to less than 80	→	$70 \leq x < 80$	8
80 or more to less than 90	→	$80 \leq x < 90$	5
90 or more to less than 100	→	$90 \leq x < 100$	2
			19

The stem-and-leaf process can be used to construct a frequency distribution; however, the stem representation is not compatible with all class widths. For example, class widths of 3, 4, and 7 are awkward to use. Thus, sometimes it is advantageous to have a separate procedure for constructing a grouped frequency distribution.

Constructing Grouped Frequency Distribution

To illustrate this grouping (or classifying) procedure, let's use a sample of 50 final exam scores taken from last semester's elementary statistics class. Table 2.5 lists the 50 scores.

Table 2.5 Statistics Exam Scores

60	47	82	95	88	72	67	66	68	98
90	77	86	58	64	95	74	72	88	74
77	39	90	63	68	97	70	64	70	70
58	78	89	44	55	85	82	83	72	77
72	86	50	94	92	80	91	75	76	78

Procedure

(1) Identify the high score ($H = 98$) and the low score ($L = 39$), and find the range:

$$\text{range} = H - L = 98 - 39 = 59$$

(2) Select a number of classes ($m = 7$) and a class width ($c = 10$) so that the product ($mc = 70$)

is a bit larger than the range (range = 59).

(3) Pick a starting point. This starting point should be a little smaller than the lowest score L . Suppose we start at 35; counting from there by tens (the class width), we get 35, 45, 55, 65, \dots , 95, 105. These are called the **class boundaries**. The classes for the data in Table 2.6 are:

Table 2.6 The classes for the data

35 or more to less than 45	\rightarrow	$35 \leq x < 45$
45 or more to less then 55	\rightarrow	$45 \leq x < 55$
55 or more to less than 65	\rightarrow	$55 \leq x < 65$
65 or more to less than 75	\rightarrow	$65 \leq x < 75$
\vdots	\vdots	$75 \leq x < 85$
		$85 \leq x < 95$
95 or more to and including 105	\rightarrow	$95 \leq x < 105$

Notes

1. At a glance you can check the number pattern to determine whether the arithmetic used to form the classes was correct (35, 45, 55, \dots , 105).
2. For the interval $35 \leq x < 45$, the 35 is the lower class boundary and 45 is the upper class boundary. Observations that fall on the lower class boundary stay in that interval; observations that fall on the upper class boundary go into the next higher interval.
3. The class width is the difference between the upper and lower class boundaries.
4. Many combinations of class widths, numbers of classes, and starting points are possible when classifying data. There is no one best choice. Try a few different combinations, and use good judgment to decide on the one to use.

◇ Basic Guidelines ◇

For Constructing a Grouped Frequency Distribution:

1. Each class should be of the same width.
2. Classes (sometimes called bins) should be set up so that they do not overlap and so that each data value belongs to exactly one class.
3. For the examples and exercises associated with this textbook, 5 to 12 classes are most desirable, since all samples contain fewer than 125 data values. (The square root of n is a reasonable guideline for the number of classes with samples of fewer than 125 data.)
4. Use a system that takes advantage of a number pattern to guarantee accuracy. (This is demonstrated below.)
5. When it is convenient, an even-numbered class width is often advantageous.

Once the classes are set up, we need to sort the data into those classes. The method used to sort will depend on the current format of the data: If the data are ranked, the frequencies can be counted; if the data are not ranked, we will tally the data to find the frequency numbers. When classifying data, it helps to use a standard chart (see Table 2.7).

Table 2.7 Standard Chart for Frequency Distribution

Class Number	Class Tallies	Boundaries	Frequency
1		$35 \leq x < 45$	2
2		$45 \leq x < 55$	2
3		$55 \leq x < 65$	7
4		$65 \leq x < 75$	13
5		$75 \leq x < 85$	11
6		$85 \leq x < 95$	11
7		$95 \leq x < 105$	4
			50

Notes

1. If the data have been ranked (list form, dotplot, or stem-and-leaf), tallying is unnecessary; just count the data that belong to each class.
2. If the data are not ranked, be careful as you tally.
3. The frequency f for each class is the number of pieces of data that belong in that class.
4. The sum of the frequencies should equal the number of pieces of data $n(n = \sum f)$. This summation serves as a good check.

Now you can see why it is helpful to have an even class width. An odd class width would have resulted in a class midpoint with an extra digit. For example, the class 45-54 is 9 wide and the class midpoint is 49.5.

Each class needs a single numerical value to represent all the data values that fall into that class. The class midpoint (sometimes called the class *mark*) is the numerical value that is exactly in the middle of each class. It is found by adding the class boundaries and dividing by 2. Table 2.8 shows an additional column for the class midpoint, x . As a check of your arithmetic, successive class midpoints should be a class width apart, which is 10 in this example (40, 50, 60, \dots , 100 is a recognizable pattern).

Table 2.8 Frequency Distribution with Class Midpoints

Class Number	Class Boundaries	Frequency f	Class Midpoints x
1	$35 \leq x < 45$	2	40
2	$45 \leq x < 55$	2	50
3	$55 \leq x < 65$	7	60
4	$65 \leq x < 75$	13	70
5	$75 \leq x < 85$	11	80
6	$85 \leq x < 95$	11	90
7	$95 \leq x < 105$	4	100
		50	

When we classify data into classes, we lose some information. Only when we have all the raw data do we know the exact values that were actually observed for each class. For example, we put a 47 and a 50 into class 2, with class boundaries of 45 and 55. Once they are placed in the class, their values are lost to us and we use the class midpoint, 50, as their representative value.

2.2.2 Histograms

One way statisticians visually represent frequency counts of a quantitative variable is to use a bar graph called a histogram. A histogram is made up of three components:

- (1) A title, which identifies the population or sample of concern.
- (2) A vertical scale, which identifies the frequencies in the various classes.
- (3) A horizontal scale, which identifies the variable x . Values for the class boundaries or class midpoints may be labeled along the x -axis. Use whichever method of labeling the axis best presents the variable.

Definition 9

■ **Class midpoint (class mark):** The numerical value that is exactly in the middle of each class.

Definition 10

■ **Histogram:** A bar graph that represents a frequency distribution of a quantitative variable.

The frequency distribution from Table 2.8 appears in histogram form in Figure 2.7.

Sometimes the **relative frequency** of a value is important. The relative frequency is a proportional measure of the frequency for an occurrence. It is found by dividing the class frequency by the total number of observations. Relative frequency can be expressed as a common fraction, in decimal form, or as a percentage. In our example about the exam scores, the frequency associated with the third class (55-65) is 7. The relative frequency for the third class is $\frac{7}{50}$, or 0.14, or 14%. Relative frequencies are often useful in a presentation because nearly everybody understands fractional parts when they are expressed as percentages. Relative frequencies are particularly useful when comparing the frequency distributions of two different size sets of data. Figure 2.8 is a relative frequency histogram of the sample of the 50 final exam scores from Table 2.8.

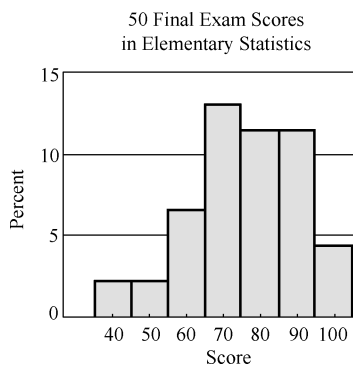


Figure 2.7 Frequency histogram

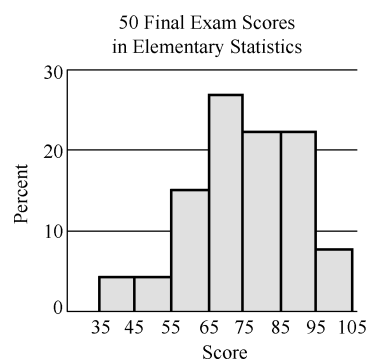


Figure 2.8 Relative frequency histogram

A stem-and-leaf display contains all the information needed to create a histogram, for example, Figure 2.5B. In Figure 2.9A the stem-and-leaf has been rotated 90° and labels have been added to show its relationship to a histogram. Figure 2.9B shows the same set of data as a completed histogram.

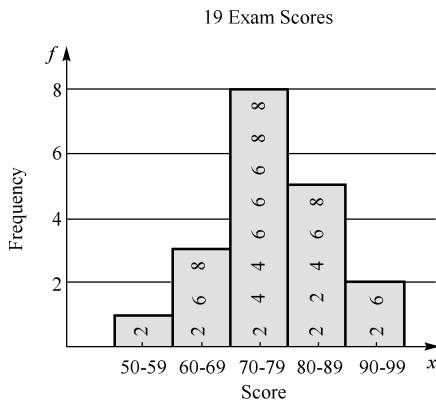


Figure 2.9A Modified stem-and-leaf display

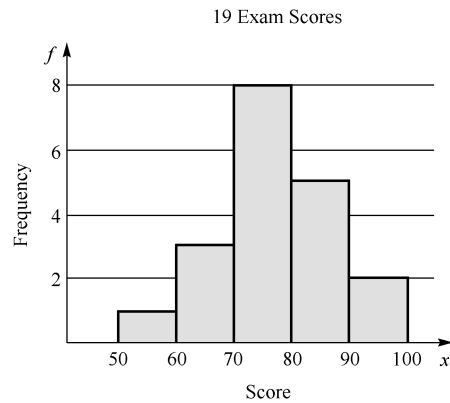


Figure 2.9B Histogram

Histograms are valuable tools. For example, the histogram of a sample should have a distribution shape very similar to that of the population from which the sample was drawn. If the reader of a histogram is at all familiar with the variable involved, he or she will usually be able to interpret several important facts. Figure 2.10 presents histograms with descriptive labels resulting from their geometric shape.

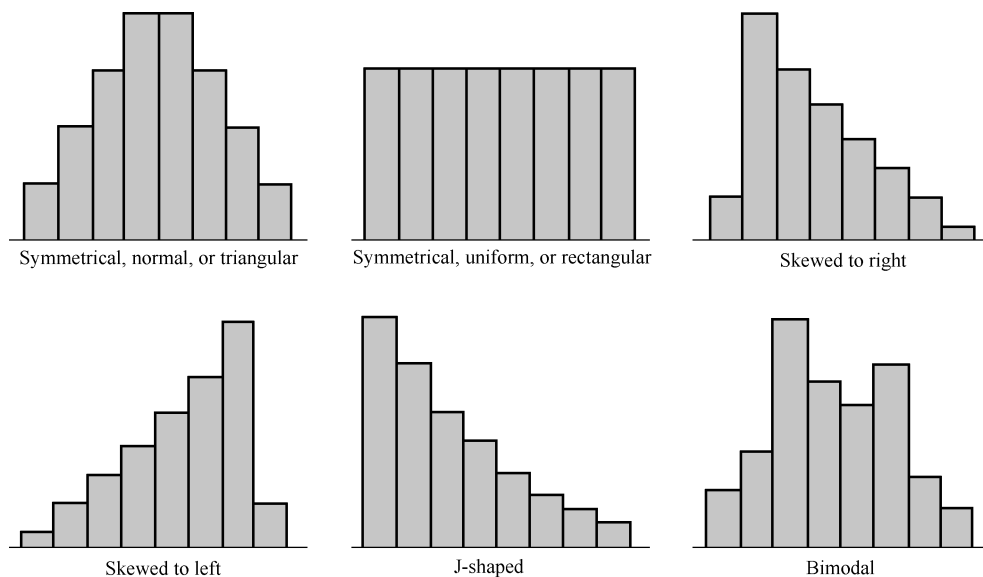


Figure 2.10 Shapes of histograms

Briefly, the terms used to describe histograms are as follows, see Figure 2.11:

Symmetrical: Both sides of this distribution are identical (halves are mirror images).

Normal (triangular): A symmetrical distribution is mounded up about the mean and becomes sparse at the extremes. (Additional properties are discussed later.)

Uniform (rectangular): Every value appears with equal frequency.

Skewed: One tail is stretched out longer than the other. The direction of skewness is on the side of the longer tail.

J-shaped: There is no tail on the side of the class with the highest frequency.

Bimodal: The two most populous classes are separated by one or more classes. This situation often implies that two populations are being sampled.

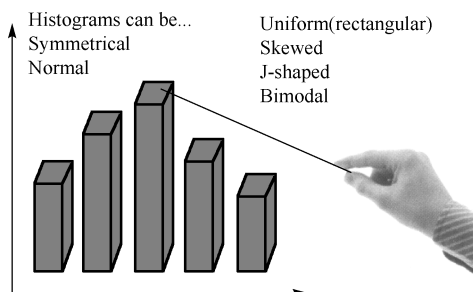


Figure 2.11 Several types of histograms

Notes

1. The mode is the value of the data that occurs with the greatest frequency. (Mode will be discussed in Unit 2.3.)
2. The modal class is the class with the highest frequency.
3. A bimodal distribution has two high-frequency classes separated by classes with lower frequencies. It is not necessary for the two high frequencies to be the same.

2.2.3 Cumulative Frequency Distribution and Ogives

Another way to express a frequency distribution is to use a **cumulative frequency distribution** to pair cumulative frequencies with values of the variable.

The cumulative frequency for any given class is the sum of the frequency for that class and the frequencies of all classes of smaller values. Table 2.9 shows the cumulative frequency distribution from Table 2.8.

Table 2.9 Using Frequency Distribution to Form a Cumulative Frequency Distribution

Class Number	Class Boundaries	Frequency f	Cumulative Frequency
1	$35 \leq x < 45$	2	2 (2)
2	$45 \leq x < 55$	2	4 (2+2)
3	$55 \leq x < 65$	7	11 (7+4)
4	$65 \leq x < 75$	13	24 (13 + 11)
5	$75 \leq x < 85$	11	35 (11 + 24)
6	$85 \leq x < 95$	11	46 (11 + 35)
7	$95 \leq x < 105$	4	50 (4 + 46)
		50	

The same information can be presented by using a *cumulative relative frequency distribution* (see Table 2.10). This combines the cumulative frequency and the relative frequency ideas.

Definition 11

- **Cumulative frequency distribution:** A frequency distribution that pairs cumulative frequencies with values of the variable.

Cumulative distributions can be displayed graphically using an ogive. Whereas a histogram is a bar graph, an ogive is a line graph of a cumulative frequency or cumulative relative frequency distribution. An ogive has the following three components:

- (1) A title, which identifies the population or sample.
- (2) A vertical scale, which identifies either the cumulative frequencies or the cumulative relative frequencies. Figure 2.12 shows an ogive with cumulative relative frequencies.
- (3) A horizontal scale, which identifies the upper class boundaries. Until the upper boundary of a class has been reached, you cannot be sure you have accumulated all the data in that class. Therefore, the horizontal scale for an ogive is always based on the upper class boundaries.

Definition 12

■ **Ogive:** A line graph of a cumulative frequency or cumulative relative frequency distribution.

Table 2.10 Cumulative Relative Frequency Distribution

Class Number	Class Boundaries	Cumulative Relative Frequency	Cumulative frequencies are for the interval 35 up to the upper boundary of that class.
1	$35 \leq x < 45$	$2/50$, or 0.04	← from 35 up to less than 45
2	$45 \leq x < 55$	$4/50$, or 0.08	← from 35 up to less than 55
3	$55 \leq x < 65$	$11/50$, or 0.22	← from 35 up to less than 65
4	$65 \leq x < 75$	$24/50$, or 0.48	
5	$75 \leq x < 85$	$35/50$, or 0.70	
6	$85 \leq x < 95$	$46/50$, or 0.92	
7	$95 \leq x < 105$	$50/50$, or 1.00	← from 35 up to less than 105

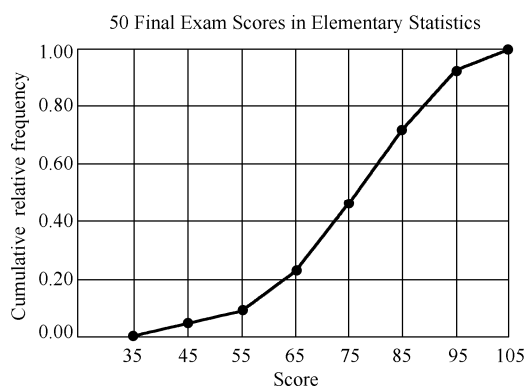


Figure 2.12 Ogive

Note: Every ogive starts on the left with a relative frequency of zero at the lower class boundary of the first class and ends on the right with a cumulative relative frequency of 100% at the upper class boundary of the last class.

New Words and Expressions

condense [kən'dens] v. 精简; 压缩; (使) 凝结

chart [tʃɑ:t] n. 图表; 航海图; 排行榜

awkward ['ɔ:kwəd] *adj.* 不方便的；笨拙的；令人尴尬的；难对付的
 mark [mɑ:k] *n.* 斑点；记号；成绩；标准
 histogram ['hɪstəgræm] *n.* 直方图，柱形图
 tally ['tæli] *n.* 计数器；标签；记账 *vt.* 测量，计数；通过做记号记录；加标签于
 midpoint ['mɪdpɔɪnt] *n.* 中点，中值，正中央 *class midpoint* 组中点（值）
 symmetrical [sɪ'metrɪkl] *adj.* 对称的，匀称的 *symmetric axis* 对称轴
 mirror ['mɪrə(r)] *n.* 镜子，反光镜；真实的写照；反映，借鉴
 bimodal [baɪ'məʊdl] *adj.* 双峰的
 skewness [sk'ju:nes] *n.* 偏斜度，偏斜；偏态
 ogive ['ɒdʒaɪv] *n.* 交错骨，尖顶拱；[统]累积曲线

Technical Terms

frequency distribution 频数分布
 class boundaries 组限，组界限
 representative value 代表值
 cumulative relative frequency 累积相对频率
 cumulative relative frequency distribution 累积相对频率分布

2.3 Measures of Central Tendency

Measures of central tendency are numerical values that location, in some sense, the center of set of data.

The term average is often associated with all measures of central tendency, including the mean, median, mode, midrange.

2.3.1 Finding the Mean

The mean, also called the arithmetic mean, is the average with which you are probably most familiar. The sample mean is represented by \bar{x} (read “ x -bar” or “sample mean”). The mean is found by adding all the values of the variable x (this sum of x values is symbolized $\sum x$) and dividing the sum by the number of these values, n (the “sample size”). We express this in formula form as

$$\text{sample mean: } x\text{-bar} = \frac{\text{sum of all } x}{\text{number of } x} = \bar{x} = \frac{\sum x}{n} \quad (2.1)$$

Note: that the population mean, μ (lowercase mu, Greek alphabet), is the mean of all x values for the entire population.

FYI: the mean is the middle point by weight.

Let's work on finding the mean using a set of data consisting of the five values 6, 3, 8, 6, and 4. To find the mean, we'll first use formula (2.1). Doing that, we find

$$\bar{x} = \frac{\sum x}{n} = \frac{6+3+8+6+4}{5} = \frac{27}{5} = 5.4$$

A physical representation of the mean can be constructed by thinking of a number line balanced on a fulcrum. A weight is placed on a number on the line corresponding to each number in the sample of our example above. In Figure 2.13 there is one weight each on the 3, 8, and 4 and two weights on the 6, since there are two 6s in the sample. The mean is the value that balances the weights on the number line in this case, 5.4.

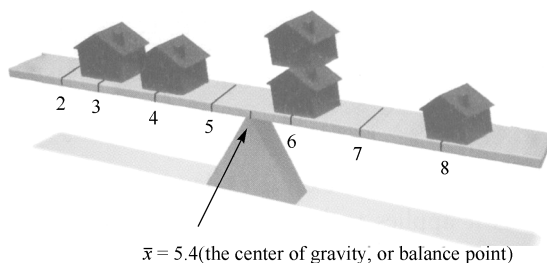


Figure 2.13 Physical representation of the mean

2.3.2 Finding the Median

The value of the data that occupies the middle position when the data are ranked in order according to size is called the median. The sample median is represented by \tilde{x} (read “x-tilde” or “sample median”). The population median, M (uppercase mu in the Greek alphabet), is the data value in the middle position of the entire ranked population.

Finding the median involves three basic steps. First, you need to rank the data. Then you determine the depth of the median. The depth (number of positions from either end), or position, of the median is determined by the formula

$$\begin{aligned} \text{depth of median} &= \frac{\text{number} + 1}{2} \\ d(\tilde{x}) &= \frac{n + 1}{2} \end{aligned} \tag{2.2}$$

The median’s depth (or position) is found by adding the position numbers of the smallest data (1) and the largest data (n) and dividing the sum by 2 (n is the number of pieces of data). Finally, you must determine the value of the median. To do this, you count the ranked data, locating the data in the $d(\tilde{x})$ th position. The median will be the same regardless of which end of the ranked data (high or low) you count from. In fact, counting from both ends will serve as an excellent check.

The following two examples demonstrate this procedure as it applies to both odd-numbered and even-numbered sets of data.

FYI: The value of $d(\tilde{x})$ is the depth of the median, NOT the value of the median.

Median For Odd n

Let’s practice finding the median by first working with an odd number n . We’ll find the

median for the set of data $\{6, 3, 8, 5, 3\}$. First, we rank the data. In this case, the data, ranked in order of size, are 3, 3, 5, 6, and 8. Next, we'll find the depth of the median: $d(\tilde{x}) = \frac{n+1}{2} = \frac{5+1}{2} = 3$ (the "3rd" position). We can now identify the median. The median is the third number from either end in the ranked data, or $\tilde{x} = 5$.

Notice that the median essentially separates the ranked set of data into two subsets of equal size, see Figure 2.14.

As in the above example, when n is odd, the depth of the median, $d(\tilde{x})$, will always be an integer. When n is even, however, the depth of the median, $d(\tilde{x})$, will always be a half-number, as shown next.

The median is the middle point by count.

Median of Even n

We can now compare the process we just completed with one in which we have an even number of points in our data set. Let's find the median of the sample 9, 6, 7, 9, 10, 8.

As before, we'll first rank the data by size. In this case, we have 6, 7, 8, 9, 9, and 10.

The depth of the median now is: $d(\tilde{x}) = \frac{n+1}{2} = \frac{6+1}{2} = 3.5$ (the "3.5th" position).

Finally, we can identify the median. The median is halfway between the third and fourth data values. To find the number halfway between any two values, add the two values together and divide the sum by 2. In this case, add the third value (8) and the fourth value (9) and then divide the sum (17) by 2. The median is $\tilde{x} = \frac{8+9}{2} = 8.5$, a number halfway between the "middle" two numbers, see Figure 2.15. Notice that the median again separates the ranked set of data into two subsets of equal size.

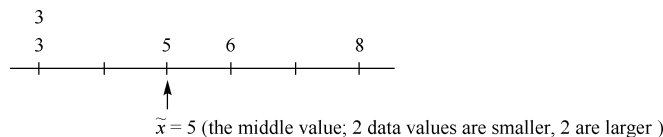


Figure 2.14 Median of $\{3, 3, 5, 6, 8\}$

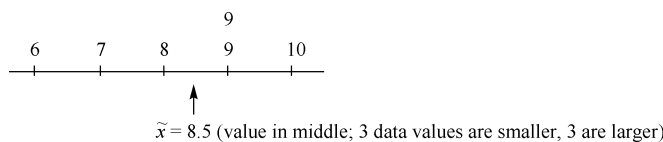


Figure 2.15 Median of $\{6, 7, 8, 9, 9, 10\}$

2.3.3 Finding the Mode

The mode is the value of x that occurs most frequently. In the set of data we used to find the median for odd n , $\{3, 3, 5, 6, 8\}$, the mode is 3, see Figure 2.16.

In the sample 6, 7, 8, 9, 9, 10, the mode is 9. In this sample, only the 9 occurs more than once;

in our earlier data set {6, 3, 8, 5, 3}, only the 3 occurs more than once. If two or more values in a sample are tied for the highest frequency (number of occurrences), we say there is no mode. For example, in the sample 3, 3, 4, 5, 5, 7, the 3 and the 5 appear an equal number of times. There is no one value that appears most often; thus, this sample has no mode.

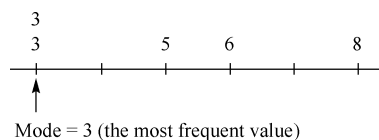


Figure 2.16 Mode of {3, 3, 5, 6, 8}

2.3.4 Finding the Midrange

The number exactly midway between a lowest data value L and a highest data value H is called the midrange. To find the midrange, average the low and the high values:

$$\text{midrange} = \frac{\text{low value} + \text{high value}}{2}$$

$$\text{midrange} = \frac{L + H}{2} \quad (2.3)$$

For the set of data {3, 3, 5, 6, 8}, $L = 3$ and $H = 8$, see Figure 2.17.

Therefore, the midrange is

$$\text{midrange} = \frac{L + H}{2} = \frac{3 + 8}{2} = 5.5$$

The four measures of central tendency represent four different methods of describing the middle. These four values may be the same, but more likely they will be different.

For the sample data set {6, 7, 8, 9, 9, 10}, the mean \bar{x} is 8.2, the median \tilde{x} is 8.5, the mode is 9, and the midrange is 8. Their relationship to one another and to the data is shown in Figure 2.18.

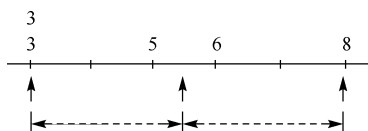


Figure 2.17 Midrange of {3, 3, 5, 6, 8}

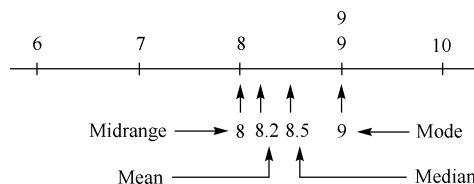


Figure 2.18 Measures of central tendency for {6, 7, 8, 9, 9, 10}

Round-off Rule

When rounding off an answer, let's agree to keep one more decimal place in our answer than was present in the original information. To avoid round-off buildup, round off only the final answer, not the intermediate steps. That is, avoid using a rounded value to do further calculations. In our previous examples, the data were composed of whole numbers; therefore, those answers that have decimal values should be rounded to the nearest tenth.

Example 2.1 "Average" Means Different Things

When it comes to convenience, few things can match that wonderful mathematical device called *averaging*. With an average you can take a fistful of figures on any subject and compute one

figure that will represent the whole fistful.

But there is one thing to remember. There are several kinds of measures ordinarily known as averages. And each gives a different picture of the figures it is called on to represent. Take an example. Table 2.11 contains the annual incomes of ten families.

Table 2.11 Annual Income of 10 Families

554,000	\$39,000	\$37,500	\$36,750	\$35,250
531,500	\$31,500	\$31,500	\$31,500	525,500

What would this group's "typical" income be? Averaging would provide the answer, so let's compute the typical income by the simpler and most frequently used kinds of averaging.

- The arithmetic mean. It is the most common form of average, obtained by adding items in the series and then dividing by the number of items: \$35,400. The mean is representative of the series in the sense that the sum of the amounts by which the higher figures exceed the mean is exactly the same as the sum of the amounts by which the lower figures fall short of the mean.
- The median. As you may have observed, six families earn less than the mean, four earn more. You might very well wish to represent this varied group by the income of the family that is right smack dab in the middle of the whole bunch. The median works out to \$33,375.
- The midrange. Another number that might be used to represent the group is the midrange, computed by calculating the figure that lies halfway between the highest and lowest incomes: \$39,750.
- The mode. So, three kinds of averages, and not one family actually has an income matching any of them. Say you want to represent the group by stating the income that occurs most frequently. That is called a mode. \$31,500 would be the modal income, see Table 2.12.

Four different averages, each valid, correct, and informative in its way. But how they differ!

Table 2.12 The four different statistic

arithmetic mean	median	midrange	mode
\$35,400	\$33,375	\$39,750	\$31,500

And they would differ still more if just one family in the group were a millionaire—or one were jobless!

So there are three lessons: First, when you see or hear an average, find out which average it is. Then you'll know what kind of picture you are being given. Second, think about the figures being averaged so you can judge whether the average used is appropriate. And third, don't assume that a literal mathematical quantification is intended every time somebody says "average". It isn't. All of us often say "the average person" with no thought of implying a mean, median, or mode. All we intend to convey is the idea of other people who are in many ways a great deal like the rest of us.

New Words and Expressions

median ['mi:diən] *n.* 中位数; 中线; [数]中值

mode [məʊd] *n.* 众数; 模式; 方式

midrange ['mɪd,rɛndʒ] *n.* 中列数; 适中范围; 中点 *adj.* 中距离的; 中程的; 中期的

midrange computer 中型电脑

fulcrum ['fʊlkrəm] *n.* 支撑杠杆的点, 支点

occupy ['ɒkjupaɪ] *vt.* 占领; 使用, 住在……; 使从事, 使忙碌

tilde ['tɪldə] *n.* 波浪符, 波浪号

odd [ɒd] *adj.* 古怪的; 奇数的; 剩余的; 临时的

even ['i:vən] *adj.* 公平的; 平坦的; 偶数的; 平均的 *vt.* 使平坦; 使相等

fistful ['fɪstfʊl] *n.* 一撮, 一把

buildup ['bɪldʌp] *n.* 组合; 集结; 累积; 形成

millionaire [ˌmɪljə'neə(r)] *n.* 百万富翁; 大富翁; 大财主

jobless ['dʒɒbləs] *adj.* 没有工作的, 失业的; 与失业有关的 *n.* 失业者 (与 the 连用)

Technical Terms

midrange 中列数

whole number 整数

arithmetic mean 算术平均数

the average person 普通人

Notes

1. 数学上关于奇偶的表示法: odd and even numbers 奇数与偶数; even and odd functions 奇函数与偶函数。

2. 同义词辨析: happen, occur, chance, take place 都可表示“发生”之意。

happen: 普通用词, 泛指一切客观事物或情况的发生, 强调动作的偶然性。

occur: 较正式用词, 可指意外地发生, 也可指意料中的发生。

chance: 侧重事前无安排或无准备而发生的事, 特指巧合。

take place: 多指通过人为安排而发生。

2.4 Measures of Dispersion

Having location the “middle” with the measures of central tendency, our search for information from data sets now turns to the measures of dispersion (spread).

The measures of dispersion include the range, variance, and standard deviation. These numerical values describe the amount of spread, or variability, that is found among the data:

Closely grouped data have relatively small values, and more widely spread out data have larger values. The closest possible grouping occurs when the data have no dispersion (all data are the same value); in this situation, the measure of dispersion will be zero. There is no limit to how widely spread out the data can be; therefore, measures of dispersion can be very large. The simplest measure of dispersion is range, which is the difference in value between the highest data value (H) and the lowest data value (L):

$$\text{range} = \text{high value} - \text{low value}$$

$$\text{range} = H - L \quad (2.4)$$

The sample 3, 3, 5, 6, 8 has a range of $H - L = 8 - 3 = 5$. The range of 5 tells us that these data all fall within a 5-unit interval, see Figure 2.19.

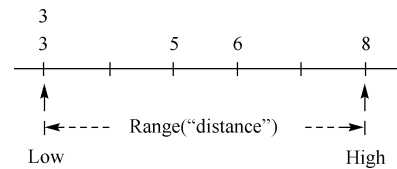


Figure 2.19 Range of {3, 3, 5, 6, 8}

The other measures of dispersion to be studied in this Unit are measures of dispersion about the mean.

To develop a measure of dispersion about the mean, let's first answer the question: How far is each x from the mean? The difference between the value of x and the mean \bar{x} , or $x - \bar{x}$, is called a deviation from the mean. Each individual value x deviates from the mean by an amount equal to $(x - \bar{x})$. This deviation $(x - \bar{x})$ is zero when x is equal to the mean \bar{x} . The deviation $(x - \bar{x})$ is positive when x is larger than \bar{x} and negative when x is smaller than \bar{x} .

When you square the deviations and take an average of those, you get something called the sample variance, s^2 . It is calculated using $n - 1$ as the divisor:

sample variance:

$$s\text{-squared} = \frac{\text{sum of (deviations squared)}}{\text{number} - 1} \quad (2.5)$$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

where n is the sample size—that is, the number of data values in the sample.

The variance of the sample 6, 3, 8, 5, 3 is calculated in Table 2.13 using formula (2.5).

Notes

1. The sum of all the x values is used to find \bar{x} .
2. The sum of the deviations, $\sum(x - \bar{x})$, is always zero, provided the exact value of 2 is used.

Use this fact as a check in your calculations, as was done in Table 2.13 (denoted by \checkmark).

Table 2.13 Calculating Variance Using Formula (2.5)

Step 1. Find $\sum x$	Step 2. Find \bar{x}	Step 3. Find each $x - \bar{x}$	Step 4. Find $\sum(x - \bar{x})^2$	Step 5. Find s^2
6	$\bar{x} = \frac{\sum x}{n}$	$6 - 5 = 1$	$(1)^2 = 1$	$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$
3		$3 - 5 = -2$	$(-2)^2 = 4$	
8		$8 - 5 = 3$	$(3)^2 = 9$	
5		$5 - 5 = 0$	$(0)^2 = 0$	

续表

Step 1. Find Σx	Step 2. Find \bar{x}	Step 3. Find each $x - \bar{x}$	Step 4. Find $\Sigma(x - \bar{x})^2$	Step 5. Find s^2
3	$\bar{x} = \frac{25}{5}$	$3 - 5 = -2$	$(-2)^2 = 4$	$s^2 = \frac{18}{4}$
$\Sigma x = 25$	$\bar{x} = 5$	$\Sigma(x - \bar{x}) = 0$	$\Sigma(x - \bar{x})^2 = 18$	$s^2 = 4.5$

3. If a rounded value of \bar{x} is used, then $\Sigma(x - \bar{x})$ will not always be exactly zero. It will, however, be reasonably close to zero.

4. The sum of the squared deviations is found by squaring each deviation and then adding the squared values.

To graphically demonstrate what variances of data sets are telling us, consider a second set of data: {1, 3, 5, 6, 10}. Note that the data values are more dispersed than the data in Table 2.12. Accordingly, its calculated variance is larger at $s^2 = 11.5$. An illustrative side-by-side graphical comparison of these two samples and their variances is shown in Figure 2.20.

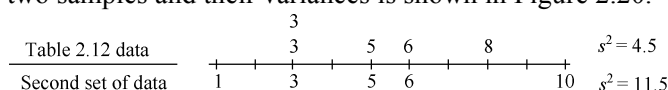


Figure 2.20 Comparison of data

2.4.1 Sample Standard Deviation

Variance is instrumental in the calculation of the standard deviation of a sample, s , which is the positive square root of the variance:

sample standard deviation:

$s = \text{square root of sample variance}$

$$s = \sqrt{s^2} \quad (2.6)$$

For the samples shown in Figure 2.20, the standard deviations are $\sqrt{4.5}$ or 2.1, and $\sqrt{11.5}$ or 3.4.

The numerator for the sample variance, $\Sigma(x - \bar{x})^2$ is often called the sum of squares for x and symbolized by $SS(x)$. Thus, formula (2.5) can be expressed as

$$\text{Sample variance:} \quad s^2 = \frac{SS(x)}{n-1} \quad (2.7)$$

The formulas for variance can be modified into other forms for easier use in various situations.

The arithmetic becomes more complicated when the mean contains nonzero digits to the right of the decimal point. However, the sum of squares for x , the numerator of formula (2.5), can be rewritten so that is not included:

$$\text{sum of squares:} \quad SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} \quad (2.8)$$

Combining formulas (2.7) and (2.8) yields the “shortcut formula” for sample variance:

$$s\text{-squared} = \frac{(\text{sum of } x^2) - \left[\frac{(\text{sum of } x)^2}{\text{number}} \right]}{\text{number} - 1}$$

$$\text{sample variance: } s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} \quad (2.9)$$

Formulas (2.8) and (2.9) are called “shortcuts” because they bypass the calculation of \bar{x} . The computations for $SS(x)$, s^2 , and s using formulas (2.8), (2.9), and (2.6) are performed as shown in Table 2.14.

Table 2.14 Calculating Standard Deviation Using the Shortcut Method

Step 1. Find $\sum x$	Step 2. Find $\sum x^2$	Step 3. Find $SS(x)$	Step 4. Find s^2	Step 5. Find s
6	$6^2=36$	$SS(x) = \sum x^2 - \frac{(\sum x)^2}{n}$	$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$	$s = \sqrt{s^2}$
3	$3^2=9$			$s = \sqrt{5.7}$
8	$8^2=64$	$SS(x) = 138 - \frac{(24)^2}{5}$		$s^2 = \frac{22.8}{4}$
5	$5^2=25$			
3	$2^2=4$	$SS(x) = 138 - 115.2$		
$\sum x = 24$	$\sum x^2 = 138$	$SS(x) = 22.8$	$s^2 = 5.7$	

The unit of measure for the standard deviation is the same as the unit of measure for the data. For example, if our data are in pounds, then the standard deviation s will also be in pounds. The unit of measure for variance might then be thought of as *units squared*. In our example of pounds, this would be *pounds squared*. As you can see, the unit has very little meaning.

New Words and Expressions

dispersion [dr'spɜːʃn] *n.* 散布；离差；差量；散布

spread [sprɛd] *n.* 范围；连续的一段时间 *vt. & vi.* 伸开；展开；（使）传播；（使）散布

range [reɪndʒ] *n.* 范围；射程；类别；[统]极差 *vt.* 排列；排序；把……分类

Technical Terms

decimal point 小数点，十进制小数点（actual decimal point 实际小数点）

deviation from the mean 均值离差；离均差；平均偏差

standard deviation of a sample 样本标准差

units squared 单位平方

Notes

几个常用的有关均值方面的术语如下。

deviation from mean: 均值离差，离均差

mean deviation: 平均差, 平均离差, 平均偏差, 平均偏移, 均差, 离均差
deviation mean: 偏离平均值

2.5 Measures of Position

Measures of position are used to describe the position a specific data value possesses in relation to rest of the data.

Quartiles and *percentiles* are two of the most popular measures of position. Other measures of position include midquartiles, 5-number summaries, and standard scores, or z-scores.

2.5.1 Quartiles

Quartiles are values of the variable that divide the ranked data into quarters; each set of data has three quartiles. The *first quartile*, Q_1 , is a number such that at most 25% of the data are smaller in value than Q_1 and at most 75% are larger. The *second quartile* is the median. The *third quartile*, Q_3 , is a number such that at most 75% of the data are smaller in value than Q_3 and at most 25% are larger, see Figure 2.21.

The procedure for determining the values of the quartiles is the same as that for percentiles, which are the values of the variable that divide a set of ranked data into 100 equal subsets; each set of data has 99 percentiles, see Figure 2.22. The k th percentile, P_k , is a value such that at most $k\%$ of the data are smaller in value than P_k and at most $(100 - k)\%$ of the data are larger, see Figure 2.23.

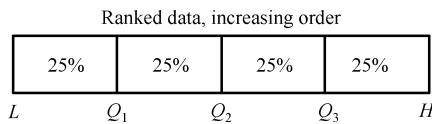


Figure 2.21 Quartiles

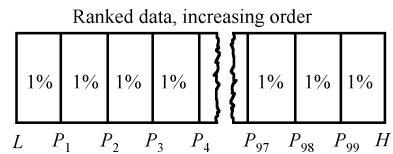


Figure 2.22 Percentiles

Notes

1. The first quartile and the 25th percentile are the same; that is, $Q_1 = P_{25}$. Also, $Q_3 = P_{75}$.

2. The median, the second quartile, and the 50th percentile are all the same: $\tilde{x} = Q_2 = P_{50}$. Therefore, when asked to find P_{50} or Q_2 , use the procedure for finding the median.

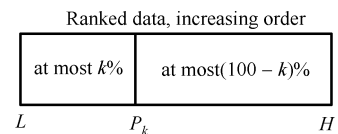


Figure 2.23 k th Percentile

Definition 13

■ **Quartiles:** Values of the variable that divide the ranked data into quarters; each set of data has three quartiles.

2.5.2 Percentiles

The procedure for determining the value of any k th percentile (or quartile) involves four basic steps as outlined on the diagram in Figure 2.24.

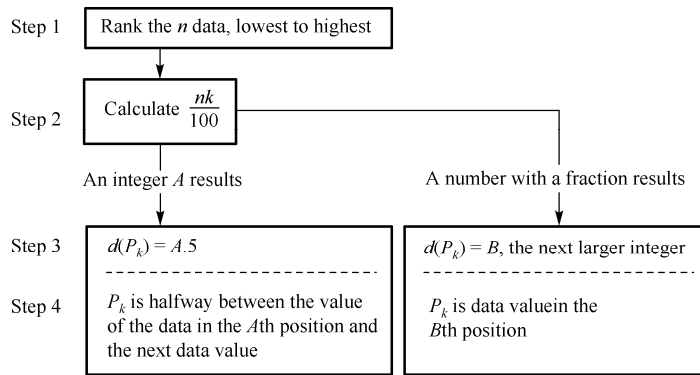


Figure 2.24 Finding P_k procedure

Using the sample of 50 elementary statistics final exam scores listed in Table 2.15, find the first quartile Q_1 , the 58th percentile P_{58} , and the third quartile Q_3 .

Table 2.15 Raw Scores for Elementary Statistics Exam

60	47	82	95	88	72	67	66	68	98	90	77	86
58	64	9.5	74	72	88	74	77	39	90	63	68	97
70	64	70	70	58	78	89	44	55	85	82	83	
72	77	72	86	50	94	92	80	91	75	76	78	

SOLUTION

Step 1 Rank the data: A ranked list may be formulated (see Table 2.15), or a graphic display showing the ranked data may be used. The dotplot and the stem-and-leaf are handy for this purpose.

Definition 14

■ **Percentiles:** Values of the variable that divide a set of ranked data into 100 equal subsets; each set of data has 99 percentiles.

The stem-and-leaf is especially helpful, since it gives depth numbers counted from both extremes when it is computer generated, see Figure 2.25. Step 1 is the same for all three statistics.

Table 2.16 Ranked Data: Exam Scores

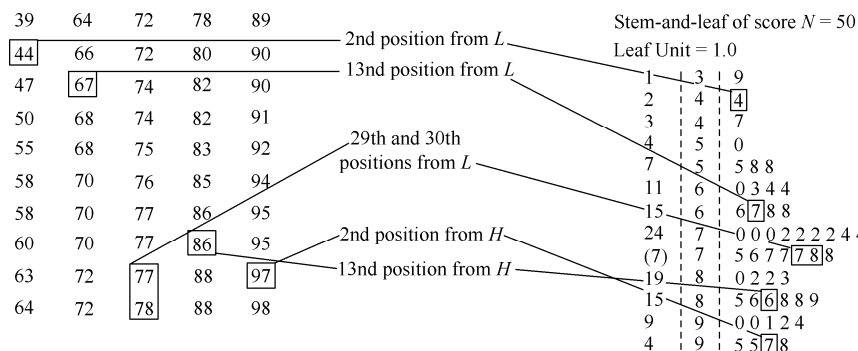


Figure 2.25 Final exam scores

Find Q_1 :

Step 2 Find $\frac{nk}{100} : \frac{nk}{100} = \frac{(50)(25)}{100} = 12.5$

Here $n = 50$ and $k = 25$, since $Q_1 = P_{25}$.

Step 3 Find the depth of Q_1 : $d(Q_1) = 13$

Since 12.5 contains a fraction, B is the next larger integer, 13.

Step 4 Find Q_1 : Q_1 is the 13th value, counting from L , see Table 2.16 or Figure 2.25, $Q_1 = 67$

Find P_{58} :

Step 2 Find $\frac{nk}{100} : \frac{nk}{100} = \frac{(50)(58)}{100} = 29$

Here $n = 50$ and $k = 58$ for P_{58} .

Step 3 Find the depth of P_{58} : $d(P_{58}) = 29.5$

Since $A = 29$, an integer, add 0.5 and use 29.5.

Step 4 Find P_{58} : P_{58} is the value halfway between the values of the 29th and the 30th pieces of data, counting from L , see Table 2.16 or Figure 2.25, so

$$P_{58} = \frac{77 + 78}{2} = 77.5$$

Therefore, it can be stated that “at most 58% of the exam grades are smaller in value than 77.5.” This is also equivalent to stating that “at most 42% of the exam grades are larger in value than 77.5.”

Optional technique: When k is greater than 150, subtract k from 100 and use $(100 - k)$ in place of k in Step 2. The depth is then counted from the largest data value H .

Find Q_3 using the optional technique:

Step 2 Find $\frac{nk}{100} : \frac{nk}{100} = \frac{(50)(25)}{100} = 12.5$

($n = 50$ and $k = 75$, since $Q_3 = P_{75}$, and $k > 50$; use $100 - k = 100 - 75 = 25$.)

Step 3 Find the depth of Q_3 from H : $d(Q_3) = 13$

Step 4 Find Q_3 : Q_3 is the 13th value, counting from H , see Table 2.16 or Figure 2.25, $Q_3 = 86$

Therefore, it can be stated that “at most 75% of the exam grades are smaller in value than 86.” This is also equivalent to stating that “at most 25% of the exam grades are larger in value than 86.”

2.5.3 Other Measures of Position

Let's now examine three other measures of position: midquartile, 5-number summary, and standard scores.

Midquartiles

Using the fundamental calculations of quartiles, you can now calculate the measure of central tendency known as the midquartile, or the numerical value midway between the first quartile and the third quartile.

$$\text{midquartile} = \frac{Q_1 + Q_3}{2} \quad (2.10)$$

So, to find the midquartile for the set of 50 exam scores given in our exam score example, you would simply add 67 to 86 and divide by 2.

$Q_1 = 67$ and $Q_3 = 86$, thus,

$$\text{midquartile} = \frac{Q_1 + Q_3}{2} = \frac{67 + 86}{2} = 76.5$$

The median, the midrange, and the midquartile are not necessarily the same value. Each is the middle value, but by different definitions of “middle”. Figure 2.26 summarizes the relationship of these three statistics as applied to our set of 50 exam scores.

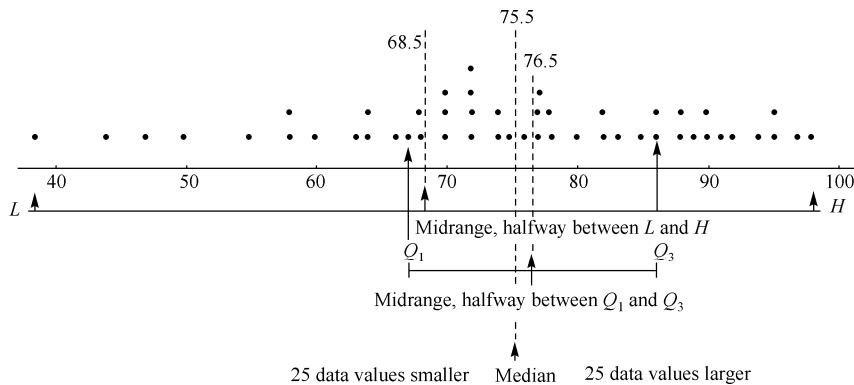


Figure 2.26 Final exam scores

Definition 15

■ **Midquartile:** The numerical value midway between the first quartile and the third quartile.

5-Number summary

Another measure of position based on quartiles and percentiles is the 5-number summary. Not only is the 5-number summary very effective in describing a set of data, it is easy information to obtain and is very informative to the reader.

The 5-number summary is composed of:

1. L , the smallest value in the data set,
2. Q_1 , the first quartile (also called P_{25} , the 25th percentile),
3. \tilde{x} , the median,
4. Q_3 , the third quartile (also called P_{75} , the 75th percentile), and
5. H , the largest value in the data set.

The 5-number summary for our set of 50 exam scores is the follow Table 2.17.

Table 2.17 The 5-number summary

39	67	75.5	86	98
L	Q_1	\tilde{x}	Q_3	H

Notice that these five numerical values divide the set of data into four subsets, with one-quarter of the data in each subset. From the 5-number summary we can observe how much the data are spread out in each of the quarters. We can now define an additional measure of dispersion. The **interquartile range** is the difference between the first and third quartiles, it is the range of the middle 50% of the data. The 5-number summary makes it very easy to see the interquartile range.

Definition 16

- **5-Number summary:** The presentation of 5 numbers that give a statistical summary of a data set: the smallest value in the data set, the first quartile, the median, the third quartile, and the largest value in the data set.

Definition 17

- **Interquartile range:** The difference between the first and third quartiles. It is the range of the middle 50% of the data.

The 5-number summary is even more informative when it is displayed on a diagram drawn to scale. A computer-generated graphic display that accomplishes this is known as the **box-and-whiskers display**. In this graphic representation of the 5-number summary, the five numerical values (smallest, first quartile, median, third quartile, and largest) are located on a scale, either vertical or horizontal. The box is used to depict the middle half of the data that lies between the two quartiles. The whiskers are line segments used to depict the other half of the data: One line segment represents the quarter of the data that is smaller in value than the first quartile, and a second line segment represents the quarter of the data that is larger in value than the third quartile.

Figure 2.27 is a box-and-whiskers display of the 50 exam scores.

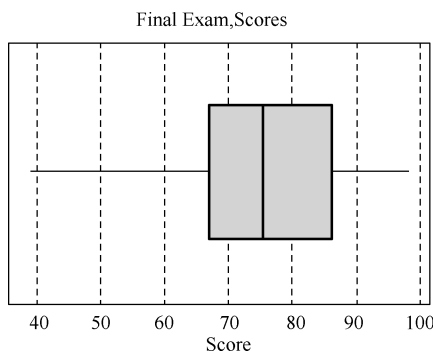


Figure 2.27 Box-and-whiskers display

Definition 18

- **Box-and-whiskers display:** A graphic representation of the 5-number summary.

Standard score (Z-scores)

So far, we've examined *general* measures of position, but sometimes it is necessary to measure the position of a *specific* value in terms of the mean and standard deviation. In those cases,

the *standard* score, commonly called the *z-score*, is used. The standard score (or *z-score*) is the position a particular value of x has relative to the mean, measured in standard deviations. The *z-score* is found by the formula:

$$z = \frac{\text{value} - \text{mean}}{\text{std.dev.}} = \frac{x - \bar{x}}{s} \quad (2.11)$$

Definition 19

■ **Standard score or *z-score*:** The position a particular value of x has relative to the mean, measured in standard deviations.

Let's apply this formula to finding the standard scores for (a) 92 and (b) 72 with respect to a sample of exam grades that has a mean score of 75.9 and a standard deviation of 11.1.

Solution

a. $x=92, \bar{x} = 75.9, s=11.1$.

Thus,
$$z = \frac{x - \bar{x}}{s} = \frac{92 - 75.9}{11.1} = \frac{16.1}{11.1} = 1.45.$$

b. $x=72, \bar{x} = 75.9, s=11.1$.

Thus,
$$z = \frac{x - \bar{x}}{s} = \frac{72 - 75.9}{11.1} = \frac{-3.9}{11.1} = -0.35.$$

This means that the score 92 is approximately one and one-half standard deviations above the mean, while the score 72 is approximately one-third of a standard deviation below the mean.

Notes

1. Typically, the calculated value of z is rounded to the nearest hundredth.
2. z -scores typically range in value from approximately -23.00 to $+13.00$.

Because the *z-score* is a measure of relative position with respect to the mean, it can be used to help us compare two raw scores that come from separate populations. For example, suppose you want to compare a grade you received on a test with a friend's grade on a comparable exam in her course. You received a raw score of 45 points; she got 72 points. Is her grade better? We need more information before we can draw a conclusion. Suppose the mean on the exam you took was 38 and the mean on her exam was 65. Your grades are both 7 points above the mean, but we still can't draw a definite conclusion. The standard deviation on the exam you took was 7 points, and it was 14 points on your friend's exam. This means that your score is one (1) standard deviation above the mean ($z = 1.0$), whereas your friend's grade is only one-half of a standard deviation above the mean ($z = 0.5$). Since your score has the "better" relative position, you conclude that your score is slightly better than your friend's score. (Again, this is speaking from a relative point of view.)

New Words and Expressions

quartile ['kwɔ:təɪl] *n.* 四分位数

approximately [ə'prɒksɪmətli] *adv.* 近似地, 大约; 大概地

one-half ['wʌnh'ɑ:f] 二分之一

Technical Terms

midquartile 中四分位数
5-number summary 五数概括法
interquartile range 四分位数间距, 四分位差
box and whiskers display 盒形图; 箱形图
line segment 线段

Notes

同义词辨析: almost, nearly, about, approximately, around, roughly 这些副词均有“大约, 差不多”之意。

almost: 指在程度上相差很小, 差不多。

nearly: nearly 与 almost 含义基本相同, 侧重指数量、时间或空间上的接近。

about: 常可与 almost 和 nearly 换用, 但 about 用于表示时间、数量的“大约”时, 实际数量可能多也可能少。

approximately: 多用于书面语, 指精确度接近某个标准以致误差可忽略不计。

around: 多用于非正式场合, 常见于美国英语。

roughly: 指粗略估计, 常代替 about。

2.6 Interpreting and Understanding Standard Deviation

Standard deviation is measure of variation (dispersion) in the data.

It has been defined as a value calculated with the use of formulas. Even so, you may be wondering what it really is and how it relates to the data. It is a kind of yardstick by which we can compare the variability of one set of data with another. This particular “measure” can be understood further by examining two statements that tell us how the standard deviation relates to the data: the *empirical rule* and *Chebyshev's theorem*.

2.6.1 The Empirical Rule and Testing for Normality

The empirical rule states that if a variable is normally distributed, then: within one standard deviation of the mean there will be approximately 68% of the data; within two standard deviations of the mean there will be approximately 95% of the data; and within three standard deviations of the mean there will be approximately 99.7% of the data. This rule applies specifically to a normal (bell-shaped) distribution, but it is frequently applied as an interpretive guide to any mounded distribution.

Figure 2.28 shows the intervals of one, two, and three standard deviations about the mean of an approximately normal distribution. Usually these proportions do not occur exactly in a sample,

but your observed values will be close when a large sample is drawn from a normally distributed population.

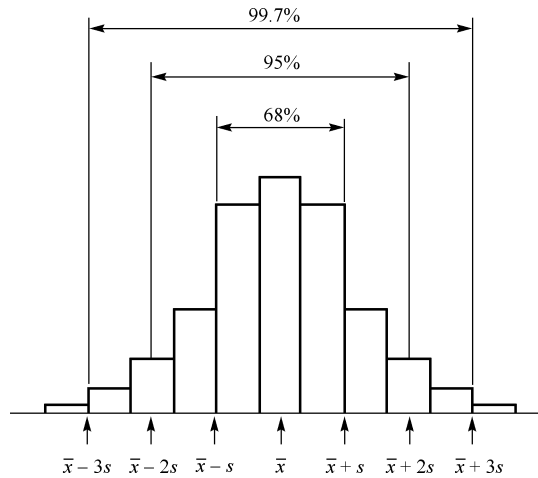


Figure 2.28 Empirical rule

If a distribution is approximately normal, it will be nearly symmetrical and the mean will divide the distribution in half (the mean and the median are the same in a symmetrical distribution). This allows us to refine the empirical rule, as shown in Figure 2.29.

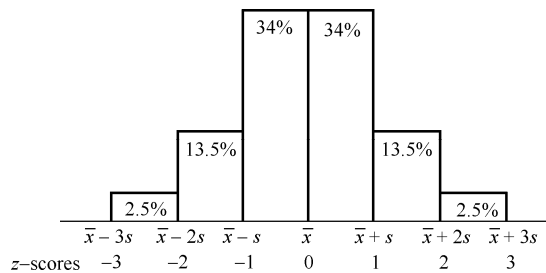


Figure 2.29 Refinement of empirical rule

◇ Empirical Rule ◇

If a variable is normally distributed, then: within one standard deviation of the mean there will be approximately 68% of the data; within two standard deviations of the mean there will be approximately 95% of the data; and within three standard deviations of the mean there will be approximately 99.7% of the data.

The empirical rule can be used to determine whether or not a set of data is approximately normally distributed. Let's demonstrate this application by working with the distribution of final exam scores that we have been using throughout this unit. The mean, \bar{x} , was found to be 75.6, and the standard deviation, s , was 14.9. The interval from one standard deviation below the mean, $\bar{x} - s$, to one standard deviation above the mean, $\bar{x} + s$, is $75.6 - 14.9 = 60.7$ to $75.6 + 14.9 = 90.5$. This interval (60.7 to 90.5) includes 61, 62, 63, ..., 89, 90. Upon inspection of the ranked data (see

Table 2.25), we see that 35 of the 50 data, or 70%, lie within one standard deviation of the mean. Furthermore, $\bar{x} - 2s = 75.6 - (2)(14.9) = 75.6 - 29.8 = 45.8$ to $\bar{x} + 2s = 75.6 + 29.8 = 105.4$ gives the interval from 45.8 to 105.4. Of the 50 data, 48, or 96%, lie within two standard deviations of the mean. All 50 data, or 100%, are included within three standard deviations of the mean (from 30.9 to 120.3). This information can be placed in a table for comparison with the values given by the empirical rule, see Table 2.18.

Table 2.18 Observed Percentages versus the Empirical Rule

Interval	Empirical Rule Percentage	Percentage Found
$\bar{x} - s$ to $\bar{x} + s$	≈ 68	70
$\bar{x} - 2s$ to $\bar{x} + 2s$	≈ 95	96
$\bar{x} - 3s$ to $\bar{x} + 3s$	≈ 99.7	100

The percentages found are reasonably close to those predicted by the empirical rule. By combining this evidence with the shape of the histogram, we can safely say that the final exam data are approximately normally distributed.

2.6.2 Chebyshev's Theorem

In the event that the data do not display an approximately normal distribution, Chebyshev's theorem gives us information about how much of the data will fall within intervals centered at the mean for all distributions. It states that the proportion of any distribution that lies within k standard deviations of the mean is at least $1 - \frac{1}{k^2}$, where k is any positive number greater than 1. This theorem applies to all distributions of data.

This theorem says that within two standard deviations of the mean ($k = 2$), you will always find at least 75% (that is, 75% or more) of the data:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4} = 0.75, \text{ at least 75\%}$$

Figure 2.30 shows a mound distribution that illustrates at least 75%.

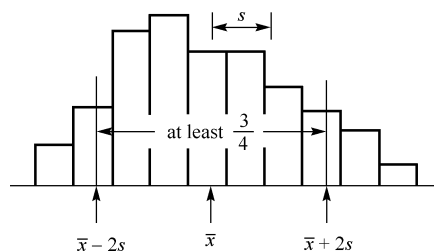


Figure 2.30 Chebyshev's theorem with $k=2$

If we consider the interval enclosed by three standard deviations on either side of the mean ($k = 3$), the theorem says that we will always find at least 89% (that is, 89% or more) of the data:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{3^2} = 1 - \frac{1}{9} = \frac{8}{9} = 0.89, \text{ at least 89\%}$$

Figure 2.31 shows a mound distribution that illustrates at least 89%.

Imagine that all the third graders at Roth Elementary School were given a physical-fitness strength test, see Table 2.19. Their test results are listed on the next page in rank order and are shown on the histogram (*data set).

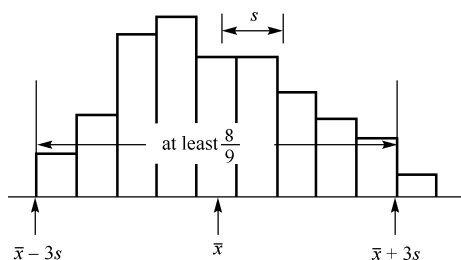


Figure 2.31 Chebyshev's theorem with $k=3$

◇ Chebyshev's Theorem ◇

The proportion of any distribution that lies within k standard deviations of the mean is at least, where k is any positive number greater than 1.

Some questions of interest are: Does this distribution satisfy the empirical rule? Does Chebyshev's theorem hold true? Is this distribution approximately normal?

Table 2.19 The Data of Some Observational Study

1	2	2	3	3	3	4	4	4	5	5	5	5	6	6	6
8	9	9	9	9	9	9	10	10	11	12	12	12	14	14	14
14	15	15	15	15	16	16	16	17	17	17	17	18	18	18	18
19	19	19	19	20	20	20	20	20	20	20	20	20	20	24	24

Histogram of Strength

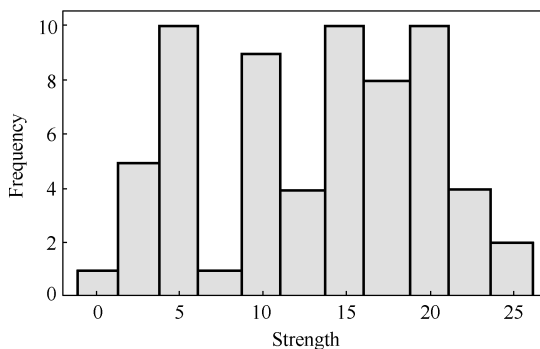


Figure 2.32 Histogram of strength and frequency

To answer the first two questions, we need to find the percentages of data in each of the three intervals about the mean. The mean is 13.0, and the standard deviation is 6.6, see Table 2.20.

It is left to you to verify the values of the mean, the standard deviation, the intervals, and the percentages.

Table 2.20 Observed Percentages versus the Empirical Rule

Mean $\pm k(\text{Std. Dev.})$	Interval	Percentage Found	Empirical	Chebyshev
13.0 \pm (6.6)	6.4 to 19.6	36/64 = 56.3%	68%	—
13.0 \pm 2(6.6)	-0.2 to 26.2	64/64 = 100%	95%	At least 75%
13.0 \pm 3(6.6)	-6.8 to 32.8	64/64 = 100%	99.70%	At least 89%

The three percentages found (56.3, 100, and 100) do not approximate the 68, 95, and 99.7 percentages stated in the empirical rule. The two percentages found (100 and 100) do agree with Chebyshev's theorem in that they are greater than 75% and 89%. Remember, Chebyshev's theorem holds for all distributions. With the distribution seen on the histogram and the three percentages found, it is reasonable to conclude that these test results are not normally distributed.

New Words and Expressions

yardstick ['jɑːdstɪk] *n.* 码尺；尺度

enclose [ɪn'kləʊz] *vt.* (用墙、篱笆等)把……围起来；把……装入信封；附入

verify ['verɪfaɪ] *vt.* 核实；证明；判定

Technical Terms

empirical rule 经验法则

Chebyshev's theorem 切比雪夫定理

approximately normally distributed 近似正态分布

Notes

同义词辨析：confirm, verify 这两个动词都有“证实”之意。

confirm：侧重以事实或以不容置疑的陈述来证实某事的正确与真实。

verify：强调以具体的事实和细节为证据。

Glossary

Mean (arithmetic mean): The mean, also called the arithmetic mean, is the average with which you are probably most familiar. The sample mean is represented by \bar{x} (read “x-bar” or “sample mean”). The mean is found by adding all the values of the variable x (this sum of x values is symbolized $\sum x$) and dividing the sum by the number of these values, n (the “sample size”).

Median: The value of the data that occupies the middle position when the data are ranked in order according to size. The sample median is represented by \tilde{x} (read “x-tilde” or “sample median”).

Mode: The mode is the value of x that occurs most frequently.

Midrange: The number exactly midway between the lowest-valued data L and the highest-valued data H .

Range: The difference in value between the highest data value (H) and the lowest data value (L).

Deviation from the mean: A deviation from the mean, $x - \bar{x}$, is the difference between the value of x and the mean \bar{x} .

Sample variance: The sample variance, s^2 , is the mean of the squared deviations.

Sample standard deviation: The standard deviation of a sample, s , is the positive square root of the variance.

Problems

2.1 The American Payroll Association got a big response to this question about company dress code: “The current dress code at my company is…”

Final results: a. A little too relaxed—27% b. A little too formal—15%
c. Just right—58%

Most people mentioned the importance of “comfort” in their explanations. The vast majority of respondents were very happy with their company’s dress code or policy.

- Construct a circle graph depicting this information. Label completely.
- Construct a bar graph depicting this same information. Label completely.
- Compare the previous two graphs, describing what you see in each one now that the graphs have been drawn and completely labeled. Do you get the same impression about these people’s feelings from both graphs? Does one emphasize anything the other one does not?

2.2 A shirt inspector at a clothing factory categorized the last 500 defects as follows: 67—missing button, 153—bad seam, 258—improperly sized, 22—fabric flaw. Construct a Pareto diagram for this information.

2.3 HoopsHype.com regularly posts the latest on the NBA. Following are the heights (in inches) of the basketball players who were the first round picks by the professional teams on June 24, 2011:

82	82	74	79	75	79	80	83	78	79
83	85	71	81	81	78	80	78	79	72
89	81	80	74	76	79	78	75	84	

- Construct a dotplot of the heights of these players.
- Use the dotplot to uncover the shortest and the tallest players,
- What is the most common height, and how many players share that height?
- What feature of the dotplot illustrates the most common height?

2.4 Construct a stem-and-leaf display of the number of points scored during each basketball game last season:

56	54	61	71	46	61	55	68
60	66	54	61	52	36	64	51

2.5 Form an ungrouped frequency distribution of the following data:

1, 2, 1, 0, 4, 2, 1, 1, 0, 1, 2, 4

Referring to the preceding distribution:

- Explain what $f = 5$ represents.
- What is the sum of the frequency column?
- What does this sum represent?

2.6 A survey of 100 resort club managers on their annual salaries resulted in the following frequency distribution:

Annual Salary (\$1000s)	15 ~ 25	25 ~ 35	35 ~ 45	45 ~ 55	55 ~ 65
No. of Managers	12	37	26	19	6

- The data value “35” belongs to which class?
- Explain the meaning of “35 ~ 45.”
- Explain what “class width” is, give its value, and describe three ways that it can be determined.
- Draw a frequency histogram of the annual salaries for resort club managers. Label class boundaries. (Retain these solutions to use in problem 2.6.)

2.7 Let’s look again at the data from problem 2.6.

Annual Salary (\$1000s)	15 ~ 25	25 ~ 35	35 ~ 45	45 ~ 55	55 ~ 65
No. of Managers	12	37	26	19	6

- Prepare a cumulative frequency distribution for the annual salaries.
- Prepare a cumulative relative frequency distribution for the annual salaries.
- Construct an ogive for the cumulative relative frequency distribution found in part b.

2.8 Consider the sample 2, 4, 7, 8, 9. Find the following:

- mean, \bar{x}
- median, \tilde{x}
- mode
- midrange

2.9 The summation $\Sigma(x - \bar{x})$ is always zero. Why? Think back to the definition of the mean (formula(2.1)) and see if you can justify this statement.

2.10 Consider the sample 2, 4, 7, 8, 9. Find the following:

- Range
- Variance s^2 , using formula (2.5)
- Standard deviation, s

2.11 Fifteen randomly selected college students were asked to state the number of hours they slept the previous night. The resulting data are 5, 6, 6, 8, 7, 7, 9, 5, 4, 8, 11, 6, 7, 8, 7. Find the following:

- Variance s^2 , using formula (2.5)
- Variance s^2 , using formula (2.9)
- Standard deviation, s

2.12 The U.S. Geological Survey collected atmospheric deposition data in the Rocky Mountains. Part of the sampling process was to determine the concentration of ammonium ions (in percentages). Here are the results from the 52 samples:

2.9	4.1	2.7	3.5	1.4	5.6	13.3	3.9	4.0
2.9	7.0	4.2	4.9	4.6	3.5	3.7	3.3	5.7

续表

3.2	4.2	4.4	6.5	3.1	5.2	2.6	2.4	5.2
4.8	4.8	3.9	3.7	2.8	4.8	2.7	4.2	2.9
2.8	3.4	4.0	4.6	3.0	2.3	4.4	3.1	5.5
4.1	4.5	4.6	4.7	3.6	2.6	4.0		

- a. Find Q_1 . b. Find Q_2 . c. Find Q_3 . d. Find the midquartile.
e. Find P_{30} . f. Find the 5-number summary. g. Draw the box-and-whiskers display.

2.13 An exam produced grades with a mean score of 74.2 and a standard deviation of 11.5.

Find the z-score for each test score x :

- a. $x = 54$ b. $x = 68$ c. $x = 79$ d. $x = 93$

2.14 A sample has a mean of 120 and a standard deviation of 20.0. Find the value of x that corresponds to each of these standard scores:

- a. $z = 0.0$ b. $z = 1.2$ c. $z = -1.4$ d. $z = 2.05$

2.15 The mean lifetime of a certain tire (轮胎) is 30,000 miles and the standard deviation is 2,500 miles.

- a. If we assume the mileages are normally distributed, approximately what percentage of all such tires will last between 22,500 and 37,500 miles?
b. If we assume nothing about the shape of the distribution, approximately what percentage of all such tires will last between 22,500 and 37,500 miles?

2.16 Using the empirical rule, determine the approximate percentage of a normal distribution that is expected to fall within the interval described.

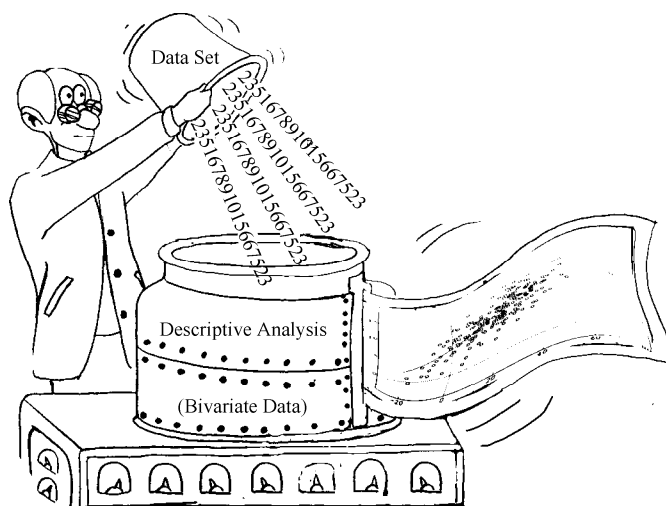
- a. Less than the mean
b. Greater than 1 standard deviation above the mean
c. Less than 1 standard deviation above the mean
d. Between 1 standard deviation below the mean and 2 standard deviations above the mean

2.17 Chebyshev's theorem guarantees that what proportion of a distribution will be included between the following.

- a. $\bar{x} - 2s$ and $\bar{x} + 2s$ b. $\bar{x} - 3s$ and $\bar{x} + 3s$

Styles in statistical analysis change over time while the object of “extracting all the information from data” or “summarization and exposure” remains the same.

— C. Radhakrishna. Rao



Unit 3

Descriptive Analysis of Bivariate Data



3.1 Bivariate Data



3.2 Linear Correlation



3.3 Linear Regression



Reading English Materials



Problems

3.1 Bivariate Data

Not all sample data can be graphically displayed with one variable. To graphically display and numerically describe sample data that involve two paired variables we need you to use **bivariate data**, which are the values of two different variables that are obtained from the same population element.

Each of the two variables may be either qualitative or quantitative. As a result, three combinations of variable types can form bivariate data:

- (i) Both variables are qualitative (attribute).
- (ii) One variable is qualitative (attribute), and the other is quantitative (numerical).
- (iii) Both variables are quantitative (both numerical).

Definition 1

■ **Bivariate data:** The values of two different variables that are obtained from the same population element.

3.1.1 Two Qualitative Variables

When bivariate data result from two qualitative (attribute or categorical) variables, the data are often arranged on a **cross-tabulation** or **contingency table**. To see how this works, let's use information on gender and college major.

Cross-Tabulation

Thirty students from our college were randomly identified and classified according to two variables: gender (M/F) and major (liberal arts, business administration, technology), as shown in Table 3.1. These 30 bivariate data can be summarized on a 2 ~ 3 cross-tabulation table, where the two rows represent the two genders, male and female, and the three columns represent the three major categories of liberal arts (LA), business administration (BA), and technology (T). The entry in each cell is found by determining how many students fit into each category. Adams is male (M) and liberal arts (LA) and is classified in the cell in the first row, first column. See the red tally mark in Table 3.2. The other 29 students are classified (tallied, shown in black) in a similar fashion.

Table 3.1 Genders and Majors of 30 College Students

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

The resulting 2×3 cross-tabulation (contingency) table, Table 3.3, shows the frequency for each cross category of the two variables along with the row and column totals, called *marginal totals* (or *marginals*). The total of the marginal totals is the *grand total* and is equal to n , the *sample size*.

Contingency tables often show percentages (relative frequencies). These percentages can be based on the entire sample or on the subsample (row or column) classifications.

Table 3.2 Cross-Tabulation of Gender and Major(tallied)

Gender	Major		
	LA	BA	T
M	(5)	(6)	(7)
F	(6)	(4)	(2)

Table 3.3 Cross-Tabulation of Gender and Major(frequencies)

Gender	Major			
	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

FYI

$$m = n \text{ (rows)}$$

$$n = n \text{ (cols)}$$

for an $m \times n$ contingency table.

Percentages Based on the Grand Total (entire sample)

The frequencies in the contingency table shown in Table 3.3 can easily be converted to percentages of the grand total by dividing each frequency by the grand total and multiplying the result by 100. For example, 6 becomes 20%:

$$\left[\left(\frac{6}{30} \right) \times 100 = 20 \right]$$

From the table of percentages of the grand total (see Table 3.4 on the next page), we can easily see that 60% of the sample were male, 40% were female, 30% were technology majors, and so on. These same statistics (numerical values describing sample results) can be shown in a bar graph, see Figure 3.1.

Table 3.4 Cross-Tabulation of Gender and Major (relative frequencies; % of grand total)

Gender	LA	BA	T	Row Total
M	17%	20%	23%	60%
F	20%	13%	7%	40%
Col. Total	37%	33%	30%	100%

Table 3.4 and Figure 3.1 show the distribution of male liberal arts students, female liberal arts students, male business administration students, and so on, relative to the entire sample.

Percentages Based on Row Totals

The frequencies in the same contingency table, Table 3.3, can be expressed as percentages of

the row totals (or gender) by dividing each row entry by that row's total and multiplying the results by 100. Table 3.5 is based on row totals. From Table 3.5 we see that 28% of the male students were majoring in liberal arts, whereas 50% of the female students were majoring in liberal arts.

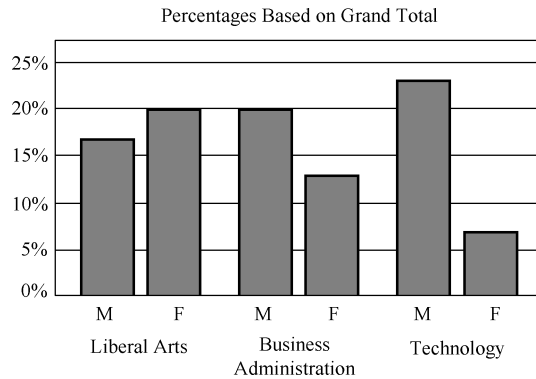


Figure 3.1 Bar graph

Table 3.5 Cross-Tabulation of Gender and Major(% of row totals)

Gender	Major			
	LA	BA	T	Row Total
M	28%	33%	39%	100%
F	50%	33%	17%	100%
Coi. Total	37%	33%	30%	100%

Percentages Based on Column Totals

The frequencies in the contingency table, Table 3.3, can be expressed as percentages of the column totals (or major) by dividing each column entry by that column's total and multiplying the result by 100. Table 3.6 is based on column totals. From Table 3.6 we see that 45% of the liberal arts students were male, whereas 55 % of the liberal arts students were female.

Table 3.6 Cross-Tabulation of Gender and Major(% of column totals)

Gender	Major			
	LA	BA	T	Row Total
M	45%	60%	78%	60%
F	55%	40%	22%	40%
Col. Total	100%	100%	100%	100%

3.1.2 One Qualitative and One Quantitative Variable

When bivariate data result from one qualitative and one quantitative variable, the quantitative values are viewed as separate samples, each set identified by levels of the qualitative variable. Each sample is described using the techniques from Unit 2, and the results are displayed side by side for easy comparison.

To see how a side-by-side comparison works, let's use the example of stopping distance. The distance required to stop a 3000-pound automobile on wet pavement was measured to compare the

stopping capabilities of three tire tread designs, see Table 3.7. Tires of each design were tested repeatedly on the same automobile on a controlled patch of wet pavement.

Table 3.7 Stopping Distances (in Feet) for Three Tread Designs

Design A($n=6$)			Design B($n=6$)			Design C($n=6$)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43

The design of the tread is a qualitative variable with three levels of response, and the stopping distance is a quantitative variable. The distribution of the stopping distances for tread design A is to be compared with the distribution of stopping distances for each of the other tread designs. This comparison may be made with both numerical and graphic techniques. Some of the available options are shown in Figure 3.2, Table 3.8, and Table 3.9.

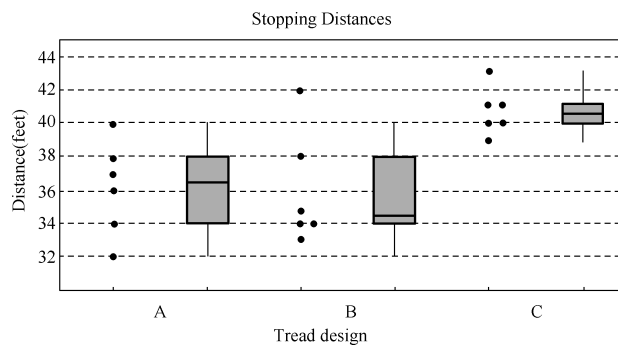


Figure 3.2 Dotplot and Box-and-Whiskers display using a common scale

Table 3.8 Five-Number Summary for Each Design

	Design A	Design B	Design C
High	40	42	43
Q_3	38	38	41
Median	36.5	34.5	40.5
Q_1	34	34	40
Low	32	33	39

Table 3.9 Mean and Standard Deviation for Each Design

	Design A	Design B	Design C
Mean	36.2	36.0	40.7
Standard deviation	2.9	3.4	1.4

3.1.3 Two Quantitative Variables

When the bivariate data are the result of two quantitative variables, it is customary to express the data mathematically as **ordered pairs** (x, y) , where x is the input variable (sometimes called the **independent variable**) and y is the output variable (sometimes called the **dependent variable**). The data are said to be *ordered* because one value, x , is always written first. They are called *paired* because for each x value, there is a corresponding y value from the same source. For example, if x is height and y is weight, then a height and a corresponding weight are recorded for each person.

The input variable x is measured or controlled in order to predict the output variable y . Suppose some research doctors are testing a new drug by prescribing different dosages and observing the lengths of the recovery times of their patients. The researcher can control the amount of drug prescribed, so the amount of drug is referred to as x . In the case of height and weight, either variable could be treated as input and the other as output, depending on the question being asked. However, different results will be obtained from the regression analysis, depending on the choice made.

Constructing a Scatter Diagram

In problems that deal with two quantitative variables, we present the sample data pictorially on a **scatter diagram**, or a plot of all the ordered pairs of bivariate data on a coordinate axis system. On a scatter diagram, the input variable, x , is plotted on the horizontal axis, and the output variable, y , is plotted on the vertical axis.

Definition 2

■ **Scatter diagram (or scatter plot):** A plot of all the ordered pairs of bivariate data on a coordinate axis system.

To illustrate, let's work with data from Mr. Chamberlain's physical fitness course in which several fitness scores were taken. The following sample contains the numbers of push-ups and sit-ups done by 10 randomly selected students:

(27, 30) (22, 26) (15, 25)(35, 42)(30, 38)
(52, 40) (35, 32) (55, 54) (40, 50)(40, 43)

Table 3.10 shows these sample data, and Figure 3.3 shows a scatter diagram of the data.

The scatter diagram from Mr. Chamberlain's physical fitness course shows a definite pattern. Note that as the number of push-ups increased so did the number of sit-ups.

Table 3.10 Data for Push-ups and Sit-ups

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	18	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

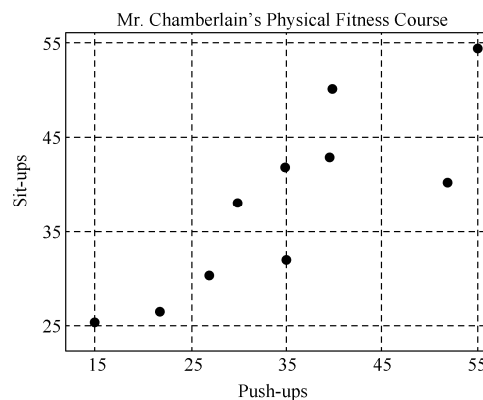


Figure 3.3 Scatter diagram

New Words and Expressions

not all 不是全部，并非所有，不是所有

major ['meɪdʒə(r)] *vi.* [美]主修(in)，专攻 *n.* 主修科目；陆军少校；成年的

column ['kɒləm] *n.* 列，纵队；圆柱；专栏

cross-tabulation [k'rɒstæbjʊl'eɪʃn] 交叉表；交叉列表；交叉制表

grand total [grænd'təʊtəl] 总共，合计；总值；累计

entry ['entri] *n.* 进入，入场；入口处，门口；登记，记录；参加比赛的人

prescribe [pri'skraɪb] *vt. & vi.* 给……开(药)；开(处方)；规定，指定遵守

pavement ['peɪvmənt] *n.* 人行道；硬路面；铺过的路面

wet [wet] *adj.* 湿的；下雨的；(儿童)尿湿尿布的 *n.* 湿地；液体；雨天；窝囊废

patch [pætʃ] *n.* 补丁，补片；斑点；小块

push-up [puʃ ʌp] 俯卧撑，复数为 push-ups

sit-up ['sɪt,ʌp] *n.* 仰卧起坐，复数为 sit-ups

Technical Terms

tread design 轮胎花纹；胎面花纹设计

contingency table 列联表

marginal totals (or marginals) 边际总和，边缘之和

physical fitness 身体适应性；身体素质；体能

ordered pairs 有序对

independent variable 自变量，独立变量

dependent variable 因变量

scatter diagram 散点图

Notes

The doctor prepared to prescribe a receipt. 医生准备开个药方。

3.2 Linear Correlation

The primary purpose of **linear correction analysis** is to measure the strength of a linear relationship between two variables.

Let's examine some scatter diagrams that demonstrate different relationships between input, or independent variables, x , and output, or dependent variables, y . If as x increases there is no definite shift in the values of y , we say there is no correlation, or no relationship between x and y . If as x increases there is a shift in the values of y , then there is a correlation. The correlation is positive when y tends to

increase and negative when y tends to decrease. If the ordered pairs (x, y) tend to follow a straight-line path, there is a linear correlation. The preciseness of the shift in y as x increases determines the strength of the linear correlation. The scatter diagrams in Figure 3.4 demonstrate these ideas.

Perfect linear correlation occurs when all the points fall exactly along a straight line, as shown in the top two graphs of Figure 3.5. The correlation can be either positive or negative, depending on whether y increases or decreases as x increases. If the data form a straight horizontal or vertical line, there is no correlation, because one variable has no effect on the other, as shown in the bottom two graphs of Figure 3.5.

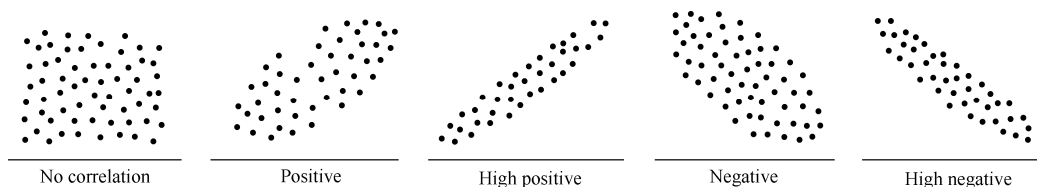


Figure 3.4 Scatter diagrams and correlation

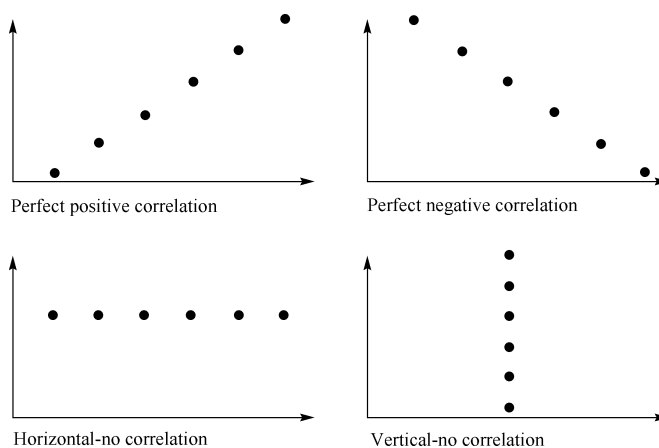


Figure 3.5 Ordered pairs forming a straight line

Scatter diagrams do not always appear in one of the forms shown in Figures 3.4 and 3.5. Sometimes they suggest relationships other than linear, as in Figure 3.6. There appears to be a definite pattern; however, the two variables are not related linearly, and therefore there is no linear correlation.

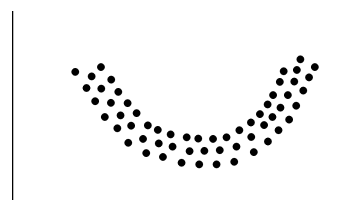


Figure 3.6 No linear correlation

3.2.1 Calculating the Linear Correlation Coefficient, r

The coefficient of linear correlation, r , is the numerical measure of the strength of the linear relationship between two variables. The coefficient reflects the consistency of the effect that a change in one variable has on the other. The value of the linear correlation coefficient helps us answer the question:

Is there a linear correlation between the two variables under consideration? The linear correlation coefficient, r , always has a value between -1 and $+1$. A value of $+1$ signifies a perfect

positive correlation, and a value of -1 shows a perfect negative correlation. If as x increases there is a general increase in the value of y , then r will be positive in value. For example, a positive value of r would be expected for the age and height of children because as children grow older, they grow taller. Also, consider the age, x , and resale value, y , of an automobile. As the car ages, its resale value decreases. Since as x increases, y decreases, the relationship results in a negative value for r . See Figure 3.7.

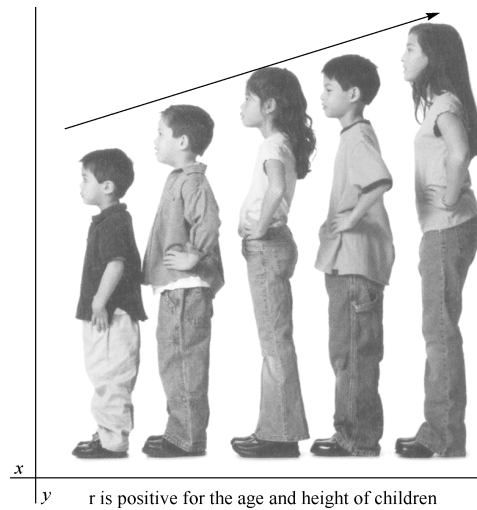


Figure 3.7 The linear correlation of x and y

The value of r is defined by Pearson's product moment formula:

Definition Formula

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n-1)s_x s_y} \quad (3.1)$$

Notes

s_x and s_y are the standard deviations of the x and y variables.

To calculate r , we will use an alternative formula, formula (3.2), that is equivalent to formula (3.1). As preliminary calculations, we will separately calculate three sums of squares and then substitute them into formula (3.2) to obtain r .

Second, to complete the preliminary calculations, we substitute the five summations (the five column totals) from the extensions table into formulas (2.8), (3.3), and (3.4), and calculate the three sums of squares:

$$\begin{aligned} SS(x) &= \sum x^2 - \frac{(\sum x)^2}{n} = 13717 - \frac{(351)^2}{10} = 1396.9 \\ SS(y) &= \sum y^2 - \frac{(\sum y)^2}{n} = 15298 - \frac{(380)^2}{10} = 850.0 \\ SS(xy) &= \sum xy - \frac{\sum x \sum y}{n} = 14257 - \frac{(351)(380)}{10} = 919.0 \end{aligned}$$

Third, we substitute the three sums of squares into formula (3.2) to find the value of the correlation coefficient:

$$r = \frac{SS(x, y)}{\sqrt{SS(x)SS(y)}} = \frac{919.0}{\sqrt{(1396.9)(858.0)}} = 0.8394 = 0.84$$

Computational Formula

$$\begin{aligned} \text{Linear correlation coefficient} &= \frac{\text{sum of squares for } xy}{\sqrt{(\text{sum of squares for } x)(\text{sum of squares for } y)}} \\ r &= \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} \end{aligned} \quad (3.2)$$

Recall the $SS(x)$ calculation from formula (2.8) for sample variance:

$$\begin{aligned} \text{sum of squares for } x &= \text{sum of } x^2 - \frac{(\text{sum of } x)^2}{n} \\ SS(x) &= \Sigma x^2 - \frac{(\Sigma x)^2}{n} \end{aligned} \quad (2.8)$$

We can also calculate:

$$\begin{aligned} \text{sum of squares for } y &= \text{sum of } y^2 - \frac{(\text{sum of } y)^2}{n} \\ SS(y) &= \Sigma y^2 - \frac{(\Sigma y)^2}{n} \end{aligned} \quad (3.3)$$

$$\begin{aligned} \text{sum of squares for } xy &= \text{sum of } xy - \frac{(\text{sum of } x)(\text{sum of } y)}{n} \\ SS(xy) &= \Sigma xy - \frac{\Sigma x \Sigma y}{n} \end{aligned} \quad (3.4)$$

So let's apply these techniques and formulae to the push-up/sit-up data from Mr. Chamberlain's fitness course.

To find the linear correlation coefficient for the push-up/sit-up data, first we construct an extensions table (Table 3.11) listing all the pairs of values (x, y) to aid us in finding x^2 , xy , and y^2 for each pair and the five column totals.

Table 3.11 Extensions Table for Finding Five Summations

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1225	42	1764	1470
5	30	900	38	1444	1140
6	52	2704	40	1600	2080
7	35	1225	32	1024	1120
8	55	3025	54	2916	2970
9	40	1600	50	2500	2000
10	40	1600	43	1849	1720
	$\Sigma x=351$	$\Sigma x^2=13717$	$\Sigma y=380$	$\Sigma y^2=15298$	$\Sigma xy=14257$
	sum of x	sum of x^2	sum of y	sum of y^2	sum of xy

Note: typically, r is rounded to the nearest hundredth.

The value of the linear correlation coefficient helps us answer the question: Is there a linear correlation between the two variables under consideration? When the calculated value of r is close to zero, we conclude that there is little or no linear correlation. As the calculated value of r changes from 0.0 toward either +1.0 or -1.0, it indicates an increasingly stronger linear correlation between the two variables. From a graphic viewpoint, when we calculate r , we are measuring how well a straight line describes the scatter diagram of ordered pairs. As the value of r changes from 0.0 toward +1.0 or -1.0, the data points create a pattern that moves closer to a straight line.

*3.2.2 Causation and Lurking Variables

As we try to explain the past, understand the present, and estimate the future, judgments about cause and effect are necessary because of our desire to impose order on our environment.

The *cause-and-effect relationship* is fairly straightforward. You may focus on the *effect* of a situation (e.g., a disease or social problem), and try to determine its *cause(s)*, or you may begin with a *cause* (unsanitary conditions or poverty) and discuss its *effect(s)*. To determine the cause of something, ask yourself **why** it happened. To determine the effect, ask yourself **what** happened. **Lurking variables** are also part of the cause-and-effect relationship being studied because even though they are not included in the study per se, they have an effect on the variables of the study and make it appear that those variables are related.

Definition 3

- **Lurking variable:** A variable that is not included in a study but has an effect on the variables of the study and makes it appear that those variables are related.

Here are some pitfalls to avoid:

1. In a direct cause-and-effect relationship, an increase (or decrease) in one variable causes an increase (or decrease) in another. Suppose there is a strong positive correlation between weight and height. Does an increase in weight *cause* an increase in height? Not necessarily. Or to put it another way, does a decrease in weight *cause* a decrease in height? Many other possible variables are involved, such as gender, age, and body type. These other variables are called *lurking variables*.
2. Some studies have shown that, a negative correlation existed between the percentage of students who received free or reduced-price lunches and the percentage of students who passed the reading proficiency test. Shall we hold back on the free lunches so that more students pass the reading test? A third variable is the motivation for this relationship, namely, poverty level.
3. Don't reason from *correlation* to *cause*: Just because all people who move to the city get old doesn't mean that the city *causes* aging. The city may be a factor, but you can't base your argument on the correlation.

Remember that a strong correlation does not necessarily imply causation.

Example 3.1 Fire Damage and Lurking Variables

The scatter plot below illustrates how the number of firefighters sent to fires (X) is related to the amount of damage caused by fires (Y) in a certain city, see Figure 3.8.

The scatter plot clearly displays a fairly strong (slightly curved) positive relationship between the two variables. Would it, then, be reasonable to conclude that sending more firefighters to a fire causes more damage, or that the city should send fewer firefighters to a fire, in order to decrease the amount of damage done by the fire? Of course not! So what is going on here?

There is a third variable in the background—the seriousness of the fire—that is responsible for the observed relationship. More serious fires require more firefighters, and also cause more damage, see Figure 3.9.

The following figure will help you visualize this situation:

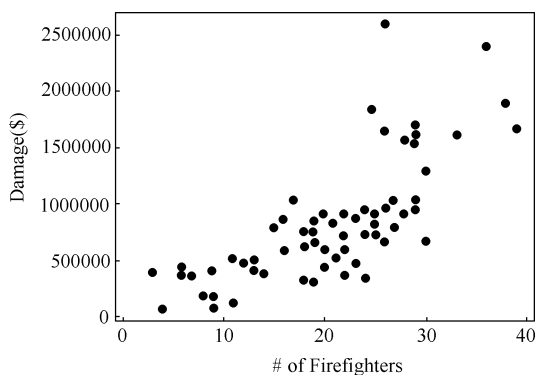


Figure 3.8 Damage and firefighters

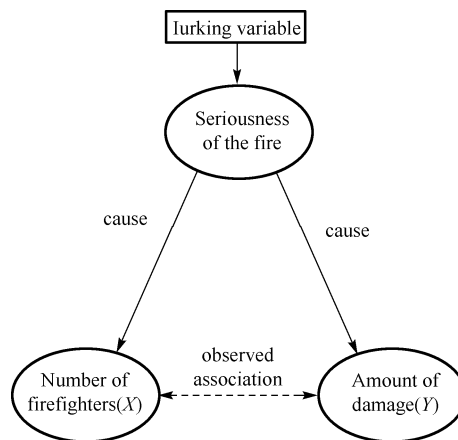


Figure 3.9 Visualization of lurking variables

Here, the seriousness of the fire is a lurking variable. A lurking variable is a variable that is not among the explanatory or response variables in a study, but could substantially affect your interpretation of the relationship among those variables.

In particular, as in our example, the lurking variable might have an effect on both the explanatory and the response variables. This common effect creates the observed association between the explanatory and response variables, even though there is no causal link between them. (We call this situation *common response*.) This possibility, that there might be a lurking variable (which we might not be thinking about) that is responsible for the observed relationship leads to our principle:

Principle: Association does not imply causation!

New Words and Expressions

preciseness [prɪ'saɪsənɪs] *n.* 准确性; 精确; 一丝不苟

coefficient [ˌkəʊfɪʃnt] *n.* 系数; (测定某种质量或变化过程的) 率; 程度

signify ['sɪgnɪfaɪ] *vt.* 表示……的意思; 意味; 预示

substitute ['sʌbstɪtju:t] *vt. & vi.* 代替, 替换, 代用 *n.* 代替者; 替补 (运动员); 替代物
impose [ɪm'pəʊz] *vt.* 强加; 征税; 以……欺骗 *vi.* 利用; 欺骗; 施加影响
per se [pɜ: 'seɪ] *adv.* 本身, 本质上
firefighter ['faɪəfaɪə(r)] *n.* 消防员, 消防队员, 消防战斗员
slightly curved 略弯的

Technical Terms

linear correction analysis 线性相关分析
perfect positive correlation 完全正相关
perfect negative correlation 完全负相关
cause-and-effect relationship 因果关系
poverty level 贫困水平线, 贫困线; 贫穷线
lurking variables 潜在变量
reduced-price lunch 低价午餐
free lunch 免费午餐
explanatory variable 解释变量
response variable 响应变量

Notes

1. under consideration 研究中, 在考虑中。类似地有, under development
2. 词根: pose ①=put, 表示“放”; ②引申为“职位”。
 - (1) compose 组成, 构成; [of]由……组成; 创作[诗歌等]
com 一起+pose 放→放到一起→组成
 - (2) depose 沉淀 de 下去+pose 放→沉淀
 - (3) discompose 解体; 使失态, 慌张
dis 分开+compose 组成, 构成→解体
 - (4) dispose [of]处理, 处置; [for]布置, 安排
dis 分开+pose 放→分开排列→安排
 - (5) expose 暴露 ex 出+pose 放→放出来→暴露
 - (6) impose 征[税]; [on]把……强加给
im 进+pose 放→进去放[强放]→强加

3.3 Linear Regression

Although the correction coefficient measures the strength of a linear relationship, it does not tell us about the mathematical relationship between the two variables.

In Unit 3.2, the correlation coefficient for the push-up/sit-up data was found to be 0.84. This

along with the pattern on the scatter diagram imply that there is a linear relationship between the number of push-ups and the number of sit-ups a student does. However, the correlation coefficient does not help us predict the number of sit-ups a person can do based on knowing he or she can do 28 push-ups. **Regression analysis** finds the equation of the line that best describes the relationship between the two variables. One use of this equation is to make predictions.

We make use of these predictions regularly—for example, predicting the success a student will have in college based on high school results and predicting the distance required to stop a car based on its speed. Generally, the exact value of y is not predictable, and we are usually satisfied if the predictions are reasonably close.

3.3.1 Line of Best Fit

If a straight-line model seems appropriate, the best-fitting straight line is found by using the **method of least squares**. Suppose that $\hat{y} = b_0 + b_1x$ is the equation of a straight line, where \hat{y} (read “y-hat”) represents the **predicted value of y** that corresponds to a particular value of x . The **least squares criterion** requires that we find the constants b_0 and b_1 such that $\Sigma(y - \hat{y})$ is as small as possible.

Figure 3.10 shows the distance of an observed value of y from a **predicted value of \hat{y}** . The length of this distance represents the value $(y - \hat{y})$ (shown as the black line segment in Figure 3.10). Note that $(y - \hat{y})$ is positive when the point (x, y) is above the line and negative when (x, y) is below the line.

Figure 3.11 shows a scatter diagram with what appears to be the **line of best fit**, along with 10 individual $(y - \hat{y})$ values. (Positive values are shown in black bold; negative, in gray.) The sum of the squares of these differences is minimized (made as small as possible) if the line is indeed the line of best fit.

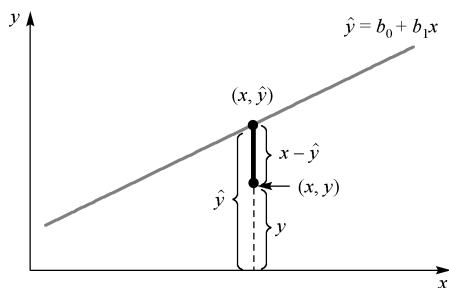


Figure 3.10 Observed and predicted values of y

Figure 3.12 shows the same data points as Figure 3.11. The 10 individual values of $(y - \hat{y})$ are plotted with a line that is definitely not the line of best fit. The value of $\Sigma(y - \hat{y})$ is 149, much larger than the 23 from Figure 3.11. Every different line drawn through this set of 10 points will result in a different value for $\Sigma(y - \hat{y})^2$. Our job is to find the one line that will make $\Sigma(y - \hat{y})^2$ the smallest possible value.

The equation of the line of best fit is determined by its **slope (b_1)** and its **y-intercept (b_0)**. The

values of the constants—slope and y-intercept—that satisfy the least squares criterion are found by using the formulas presented next.

Definition Formula

$$\text{slope: } b_1 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \quad (3.5)$$

We will use a mathematical equivalent of formula (3.5) for the slope, b_1 , that uses the sums of squares found in the preliminary calculations for correlation:

Computational Formula

$$\text{slope: } b_1 = \frac{SS(xy)}{SS(x)} \quad (3.6)$$

Notice that the numerator of formula (3.6) is the $SS(xy)$ formula (3.4) and the denominator is formula (2.8) (see previous Unit 2) from the correlation coefficient calculations. Thus, if you have previously calculated the linear correlation coefficient using the procedure outlined in previous section, you can easily find the slope of the line of best fit. If you did not previously calculate r , set up a table similar to Table 3.11 and complete the necessary preliminary calculations.

For the y-intercept, we have:

Computational Formula

$$y - \text{intercept} = \frac{(\text{sum of } y) - [(slop)(\text{sum of } x)]}{\text{number}}$$

$$b_0 = \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n} \quad (3.7)$$

Alternative Computational Formula

$$y - \text{intercept} = y - \text{bar} - (\text{slope} \cdot x - \text{bar})$$

$$b_0 = \bar{y} - (b_1 \cdot \bar{x}) \quad (3.7a)$$

Now let's reconsider the data from Mr. Chamberlain's physical class and the question of predicting a student's number of sit-ups based on the number of push-ups. We want to find the line of best fit, $\hat{y} = b_0 + b_1x$. The preliminary calculations have already been completed in Table 3.11. To calculate the slope, b_1 , using formula (3.6), recall that $SS(xy) = 919.0$ and $SS(x) = 1396.9$. Therefore,

$$\text{slope: } b_1 = \frac{SS(xy)}{SS(x)} = \frac{919.0}{1396.9} = 0.6579 = 0.66$$

To calculate the y-intercept, b_0 , using formula (3.7), recall that $\Sigma x = 351$ and $\Sigma y = 380$ from the extensions table. We have

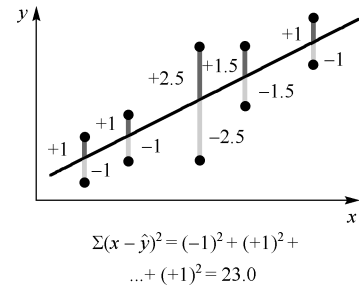


Figure 3.11 The line of best fit

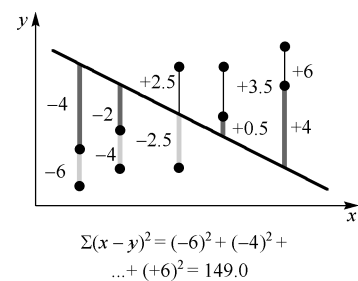


Figure 3.12 Not the Line of Best Fit

$$\begin{aligned}
 y\text{-intercept} : b_0 &= \frac{\Sigma y - (b_1 \cdot \Sigma x)}{n} = \frac{380 - 0.6579 \times 351}{10} \\
 &= \frac{380 - 230.9229}{10} = 14.9077 = 14.9
 \end{aligned}$$

By placing the two values just found into the model $\hat{y} = b_0 + b_1x$, we get the equation of the line of best fit:

$$\hat{y} = 14.9 + 0.66x$$

Notes

1. Remember to keep at least three extra decimal places while doing the calculations to ensure an accurate answer.
2. When rounding off the calculated values of b_0 and b_1 , always keep at least two significant digits in the final answer.

Now that we know the equation for the line of best fit, let's draw the line on the scatter diagram so that we can see the relationship between the line and the data. We need two points in order to draw the line on the diagram. Select two convenient x values, one near each extreme of the domain ($x = 10$ and $x = 60$ are good choices for this illustration), and find their corresponding y values.

$$\text{For } x = 10 : \hat{y} = 14.9 + 0.66x = 14.9 + 0.66(10) = 21.5; (10, 21.5)$$

$$\text{For } x = 60 : \hat{y} = 14.9 + 0.66x = 14.9 + 0.66(60) = 54.5; (60, 54.5)$$

These two points, $(10, 21.5)$ and $(60, 54.5)$, are then located on the scatter diagram (we use a purple + to distinguish them from data points) and the line of best fit is drawn (shown in black in Figure 3.13).

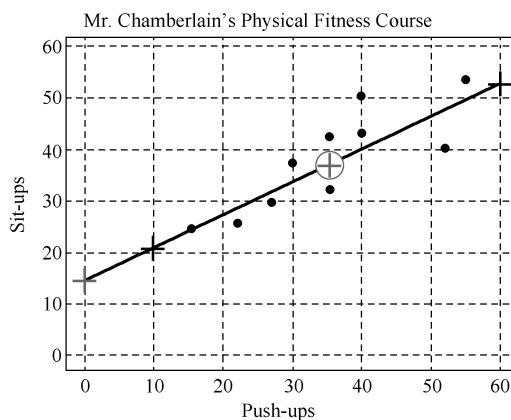


Figure 3.13 Line of best fit for push-ups versus sit-ups

There are some additional facts about the least squares method that we need to discuss.

(1) The slope, b_1 , represents the predicted change in y per unit increase in x . In our example, where $b_1 = 0.66$, if a student can do an additional 10 push-ups(x), we predict that he or she would be able to do approximately 7 (0.66×10) additional sit-ups (y).

(2) The y -intercept is the value of y where the line of best fit intersects the y -axis. (When the vertical scale is located above $x = 0$, the y -intercept is easily seen on the scatter diagram, shown as a green + in Figure 3.11.)

First, however, in interpreting b_0 , you must consider whether $x = 0$ is a realistic x value before you conclude that you would predict $\hat{y} = b_0$ if $x = 0$. To predict that if a student did no push-ups, he or she would still do approximately 15 sit-ups ($b_0 = 14.9$) is probably incorrect. Second, the x value of zero may be outside the domain of the data on which the regression line is based. In predicting y based on an x value, check to be sure that the x value is within the domain of the x values observed.

(3) The line of best fit will always pass through the *centroid*, the point (\bar{x}, \bar{y}) . When drawing the line of best fit on your scatter diagram, use this point as a check. For our illustration,

$$\bar{x} = \frac{\Sigma x}{n} = \frac{351}{10} = 35.1, \bar{y} = \frac{\Sigma y}{n} = \frac{380}{10} = 38.0$$

We see that the line of best fit does pass through $(\bar{x}, \bar{y}) = (35.1, 38.0)$, as shown \oplus in Figure 3.11.

Let's work through another example to clarify the steps involved in regression analysis.

Table 3.12 College Women's Heights and Weights

	1	2	3	4	5	6	7	8
Height, x	65	65	62	67	69	65	61	67
Weight, y	105	125	110	120	140	135	95	130

Table 3.13 Preliminary Calculations Needed to Find b_1 and b_0

	Student Height, x	x^2	Weight, y	xy
1	65	4,225	105	6,825
2	65	4,225	125	8,125
3	62	3,844	110	6,820
4	67	4,489	120	8,040
5	69	4,761	140	9,660
6	65	4,225	135	8,775
7	61	3,721	95	5,795
8	67	4,489	130	8,710
	$\Sigma x = 521$	$\Sigma x^2 = 33,979$	$\Sigma y = 960$	$\Sigma xy = 62,750$

Calculating the Line of Best Fit Equation

In a random sample of eight college women, each woman was asked her height (to the nearest inch) and her weight (to the nearest 5 pounds). The data obtained are shown in Table 3.12. Find an equation to predict the weight of a college woman based on her height (the equation of the line of best fit), and draw it on the scatter diagram in Figure 3.14.

Before we start to find the equation for the line of best fit, it is often helpful to draw the scatter diagram, which provides visual insight into the relationship between the two variables. The scatter diagram for the data on the heights and weights of college women, shown in Figure 3.14, indicates that the linear model is appropriate.

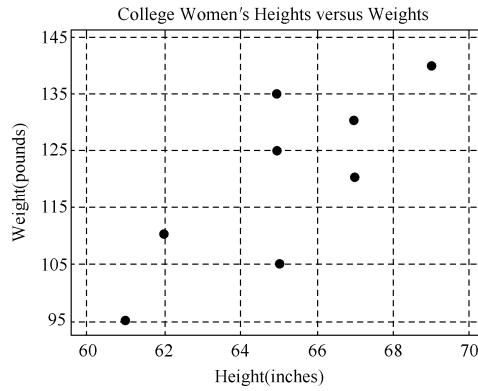


Figure 3.14 Scatter diagram

To find the equation for the line of best fit, we first need to complete the preliminary calculations, as shown in Table 3.13. The other preliminary calculations include finding $SS(x)$ from formula (2.8) and $SS(xy)$ from formula (3.4):

$$SS(x) = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 33979 - \frac{521^2}{8} = 48.875$$

$$SS(xy) = \Sigma xy - \frac{\Sigma x \Sigma y}{n} = 62750 - \frac{521 \times 960}{8} = 230.0$$

Second, we need to find the slope and the y -intercept using formulas (3.6) and (3.7):

$$\text{slope: } b_1 = \frac{SS(xy)}{SS(x)} = \frac{230.0}{48.872} = 4.706 = 7.71$$

$$y\text{-intercept: } b_0 = \frac{\Sigma y(b_1 \cdot \Sigma x)}{n} = \frac{960 - 4.706 \times 521}{8} = -186.478 = -186.5$$

Thus, the equation of the line of best fit is

$$\hat{y} = -186.5 + 4.71x$$

To draw the line of best fit on the scatter diagram, we need to locate two points. Substitute two values for x —for example, 60 and 70—into the equation for the line of best fit and obtain two corresponding values for \hat{y} :

$$\begin{aligned}\hat{y} &= -186.5 + 4.71x = -186.5 + (4.71)(60) \\ &= -186.5 + 282.6 = 96.1 \approx 96 \\ \hat{y} &= -186.5 + 4.71x = -186.5 + (4.71)(70) \\ &= -186.5 + 329.7 = 143.2 \approx 143\end{aligned}$$

The values (60, 96) and (70, 143) represent two points (designated by a black + in Figure 3.15) that enable us to draw the line of best fit.

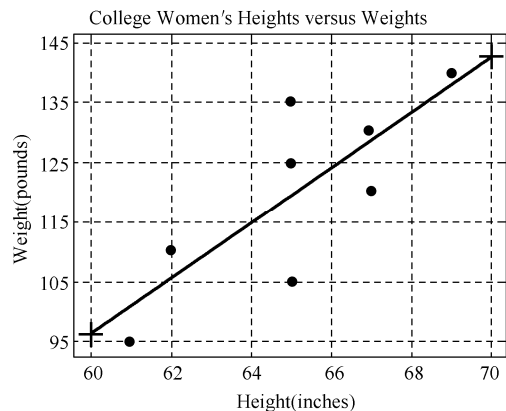


Figure 3.15 Scatter diagram with line of best fit

3.3.2 Making Predictions

One of the main reasons for finding a regression equation is to make predictions. Once a linear relationship has been established and the value of the input variable x is known, we can predict a value of y , \hat{y} . Consider the equation $\hat{y} = -186.5 + 4.71x$ relating the height and weight of college women. If a particular female college student is 66 inches tall, what do you predict her weight to be? The predicted value is

$$\hat{y} = -186.5 + 4.71x = -186.5 + 4.71 \times 66 = -186.5 + 310.86 = 124.36 \approx 124(\text{lb})$$

You should not expect this predicted value to occur exactly; rather, it is the average weight you would expect for all female college students who are 66 inches tall.

When you make predictions based on the line of best fit, observe the following restrictions:

1. The equation should be used to make predictions only about the population from which the sample was drawn. For example, using our relationship between the height and weight of college women to predict the weight of professional athletes given their height would be questionable.

2. The equation should be used only within the sample domain of the input variable. We know the data demonstrate a linear trend within the domain of the x data, but we do not know what the trend is outside this interval. Hence, predictions can be very dangerous outside the domain of the x data. For instance, in our current example, it is nonsense to predict that a college woman of height zero will weigh—186.5 pounds. Do not use a height outside the sample domain of 61 to 69 inches to predict weight. On occasion, you might wish to use the line of best fit to estimate values outside the domain interval of the sample. This can be done, but you should do it with caution and only for values close to the domain interval.

3. If the sample was taken in 2006, do not expect the results to have been valid in 1929 or to hold in 2010. The women of today may be different from the women of 1929 and the women in 2010.

New Words and Expressions

predictable [prɪ'dɪktəbl] *adj.* 可预言的; 可预报的; 可预见的;

minimize ['mɪnɪmaɪz] *vt.* [数]最小化; 求……极小值; 使减少(或缩小)到最低限度

slope [sləʊp] *n.* 斜坡; 斜面; 倾斜; 斜率

intercept [ˌɪntə'sept] *n.* [数]截距; 中途夺取, 侦听; 阻留; 定方位

clarify ['klærəfaɪ] *vt.* (尤指通过加热使黄油)纯净; 说明; 使(头脑、神智等)清醒

locate [ləʊ'keɪt] *vt.* 位于; 查找……的地点; 确定……的位置

athlete ['æθli:t] *n.* 运动员; 体育家; 强壮的人; 复数为 athletes

Technical Terms

method of least squares 最小二乘法

least squares criterion 最小二乘准则

line of best fit 最优拟合直线

significant digit 有效数字，又称 significant figure

Notes

1. 同义词辨析 explain, interpret, illustrate, clarify, account 这些动词均有“说明”之意。
explain: 含义广，最普通用词，指把某事向原来不了解、不清楚的人解释明白、说清楚等。
interpret 着重以特殊的知识、经验来解释难理解的事情。
illustrate 多指用实例或插图、图表加以说明。
clarify 指把已发生的事件、情况和现状说清楚。
account 说明某事物如何符合自然法则或逻辑。

Passage 1. The First Regression

Sir Francis Galton related the heights of sons to the heights of their fathers with a regression line. The slope of his line was less than 1. That is, sons of tall fathers were tall, but not as much above the average heights as their fathers had been above their mean. Sons of short fathers were short, but generally not as far from their mean as their fathers. Galton interpreted the slope correctly as indicating a “regression” toward the mean height—and “regression” stuck as a description of the method he had used to find the line.

Sir Francis Galton was the first to speak of “regression”, although others had fit lines to data by the same method.

Passage 2. Simpson’s Paradox

Here’s an example showing that combining percentages across very different values or groups can give confusing results. Suppose there are two sales representatives, Peter and Katrina. Peter argues that he’s the better salesperson, since he managed to close 83% of his last 120 prospects compared with Katrina’s 78%. But let’s look at the data a little more closely. Here Table 3.15 are the results for each of their last 120 sales calls, broken down by the product they were selling.

Look at the sales of the two products separately. For printer paper(打印机纸) sales, Kitrian had a 95% success rate, and Peter only had 90% rate. When selling flash drives(闪存盘), Katrina closed her sales 75% of time, but Peter only 50%. So Peter has better “overall” performance, but Katrian is better selling each product. How can this be?

		Product		
		Printer Paper	USB Flash Drive	Overall
Sales Rep	Peter	90 out of 100 90%	10 out of 20 50%	100 out of 120 83%
	Katrina	19 out of 20 95%	75 out of 100 75%	94 out of 120 78%

Table 3.15 Look at the percentages within each Product category. Who has Better success rate closing sales of paper? Who has better success rate closing sales of Flash Drives? Who has the better performance overall?

This problem is known as **Simpson’s Paradox** (辛普森悖论), name for the statistician who describes it in the 1960s. Although it is rate, there have been a few well-publicized cases of it. As we seen from the example, the problem results from inappropriately combining percentages of different groups. Katrian concentrates on selling flash drives, which is more difficult, so her *Overall*

percentage is heavily influenced but her flash drive average. Peter sells more printer paper, which appears to be easier to sell. With their different patterns of selling, taking an overall percentage is misleading. Their manager should be careful not conclude rashly that Peter is the better salesperson.

The lesson of Simpson's Paradox is to be sure combine only comparable measurements for comparable individuals. Be especially careful when combining across different levels of a second variable. It's usually better to compare percentages *within* each level, rather than across levels.

Problems

3.1 The “Outlook for Business Travelers” graphic shows two circle graphs, each with four sections. This same information could be represented in the form of a 2×4 contingency table of two qualitative variables.

- Identify the population and name the two variables.
- Construct the contingency table using entries of percentages based on row totals.

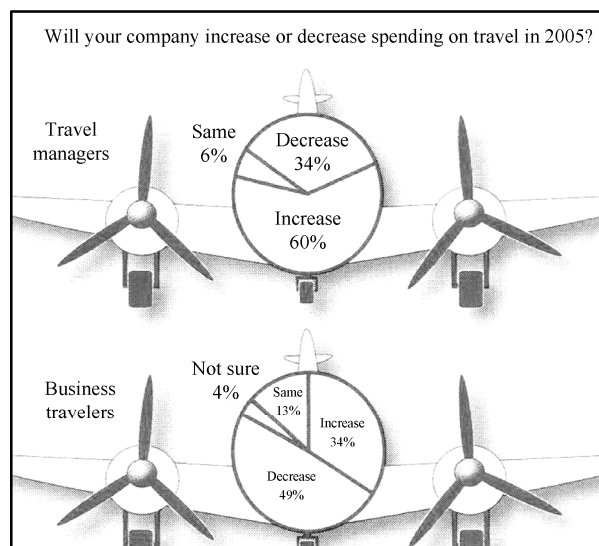


Figure 3.16 Outlook for Business travelers

3.2 Draw a coordinate axis and plot the points (0, 6), (3, 5), (3, 2), (5, 0) to form a scatter diagram. Describe the pattern that the data show in this display.

3.3 The accompanying data show the number of hours, x , studied for an exam and the grade received, y (y is measured in tens; that is, $y = 8$ means that the grade, rounded to the nearest 10 points, is 80). Draw the scatter diagram. (Retain this solution to use in problem 3.6)

x	2	3	3	4	4	5	5	6	6	6	7	7	7	8	8
y	5	5	7	5	7	7	8	6	9	8	7	9	10	8	9

3.4 Many organizations offer “special” magazine subscription rates to their members. The American Federation of Teachers is no different, and here are a few of the rates they offer their members.

Magazine	Usual Rate	Your Price
Cosmopolitan	\$29.97	\$18.00
Sports Illustrated	\$578.97	\$539.75
Ebony	\$520.00	\$514.97
Rolling Stone	\$523.94	\$511.97
Martha Stewart Living	\$524.95	\$20.00

Source: American Federation of Teachers

a. Construct a scatter diagram with “Your Price” as the dependent variable, y , and “Usual Rate” as the independent variable, x . Find

b. $SS(x)$

c. $SS(y)$

d. $SS(xy)$

e. Pearson’s product moment, r

3.5 a. Use the scatter diagram you drew in problem 3.3 to estimate r for the sample data on the number of hours studied and the exam grade.

b. Calculate r .

3.6 A marketing firm wished to determine whether the number of television commercials broadcast were linearly correlated with the sales of its product. The data, obtained from each of several cities, are shown in the following table.

City	A	B	C	D	E	F	G	H	I	J
Commercials, x	12	6	9	15	11	15	8	16	12	6
Sales Units, y	7	5	10	14	12	9	6	11	11	8

a. Draw a scatter diagram.

b. Estimate r .

c. Calculate r .

3.7 a. Use the scatter diagram you drew in problem 3.6 to estimate r for the sample data on the number of hours studied and the exam grade.

b. Calculate r .

3.8 Draw a scatter diagram for these data:

x	2	12	4	6	9	4	11	3	10	11	3	1	13	12	14	7	2	8
y	4	8	10	9	10	8	8	5	10	9	8	3	9	8	8	11	6	9

Would you be justified in using the techniques of linear regression on these data to find the line of best fit? Explain.

3.9 AJ used linear regression to help him understand his monthly telephone bill. The line of best fit was $\hat{y} = 23.65 + 1.28x$, where x is the number of long-distance calls made during a month, and y is the total telephone cost for a month. In terms of number of long-distance calls and cost:

a. Explain the meaning of the y -intercept, 23.65.

b. Explain the meaning of the slope, 1.28.

3.10 A study was conducted to investigate the relationship between the resale price, y (in hundreds of dollars), and the age, x (in years), of midsize luxury American automobiles. The equation of the line of best fit was determined to be $\hat{y} = 185.7 - 21.52x$.

a. Find the resale value of such a car when it is 3 years old.

b. Find the resale value of such a car when it is 6 years old.

c. What is the average annual decrease in the resale price of these cars?

3.11 Please translate the following discrimination content.

One famous example of Simpson's Paradox arose during an investigation of admission rates for men and women at the University of California at Berkeley's graduate schools. As reported in an article in *Science*, about 45% of male applicants were admitted, but only about 30% female applications got in. It looks like a clear case of discrimination. However, when the data were broken down by school (Engineering, Law, Medicine, etc), it turns out that within each school, the women were admitted at nearly the same or, in some cases, much higher rates than the men.

How could this be? Women applied in large numbers to schools with very low admission rates. (Law and Medicine, for example, admitted fewer than 10%.) Men tended to apply to Engineering and Science. Those schools have admission rates about 50%. When the total applicant pool was combined and the percentages were computed, the women had a much lower overall rate, but the combined percentage didn't really make sense.








The theory of probability lies at the root of all statistical theory.

—— Professor Sir John Kingman FRS, was the third director of
the Isaac Newton Institute for Mathematical Sciences,
serving from October 2001 to September 2006.



Unit 4

Introduction to Probability

-  4.1 Sample Spaces, Events and Sets
-  4.2 Probability Axioms and Simple Counting Problems
-  4.3 Permutations and Combinations
-  4.4 Conditional Probability and the Multiplication Rule
-  4.5 Independent Events, Partitions and Bayes Theorem
-  Reading English Materials
-  Problems

4.1 Sample Spaces, Events and Sets

4.1.1 Introduction

Since Statistics involves the collection and interpretation of data, we must first know how to understand, display and summarize large amounts of quantitative information, before undertaking a more sophisticated analysis. Statistical analysis of quantitative data is important throughout the pure and social sciences.

Example 4.1 Survival of cancer patients

A cancer patient wants to know the probability that he will survive for at least 5 years. By collecting data on survival rates of people in a similar situation, it is possible to obtain an empirical estimate of survival rates. We cannot know whether or not the patient will survive, or even know exactly what the *probability* of survival is. However, we can *estimate* the *proportion* of patients who survive from *data*.

Example 4.2 Car maintenance

When buying a certain type of new car, it would be useful to know how much it is going to cost to run over the first three years from new. Of course, we cannot predict exactly what this will be—it will vary from car to car. However, collecting data from people who bought similar cars will give some idea of the *distribution* of costs across the *population* of car buyers, which in turn will provide information about the *likely* cost of running the car.

4.1.2 Sample Spaces

There are lots of phenomena in nature, like tossing a coin or tossing a die, whose outcomes cannot be predicted with certainty in advance, but the set of all the possible outcomes is known. These are what we call *random phenomena* or *random experiments*. Probability theory is concerned with such random phenomena or random experiments.

Consider a random experiment. The set of all the possible outcomes is called the *sample space* of the experiment and is usually denoted by S . Any subset E of the sample space S is called an *event*. The sample space is chosen so that exactly one outcome will occur.

Suppose that E is an event. We say that the event E “occurs” if the outcome of the experiment is contained in E .

The size of the sample space is *finite*, *countably infinite* or *uncountably infinite*. Here are some examples.

These are examples of the finite sample space.

Example 4.3 Tossing a coin

The sample space is $S = \{H, T\}$. $E = \{H\}$ is an event, see Figure 4.1.



Figure 4.1 Coin is thrown

Example 4.4 Tossing a die

The sample space is $S = \{1, 2, 3, 4, 5, 6\}$. $E = \{2, 4, 6\}$ is an event, which can be described in words as “the number is even”.

Example 4.5 Tossing a coin twice

The sample space is $S = \{HH, HT, TH, TT\}$. $E = \{HH, HT\}$ is an event, which can be described in words as “the first toss results in a Heads”.

Example 4.6 Germinating seeds

The outcome of one of any replication is the number of germinating seeds. We consider 100 seed germinations, the number germinating could be anything from 0 to 100, see Figure 4.2. So, the sample space for the outcomes of this experiment is

$$S = \{0, 1, 2, \dots, 100\}.$$

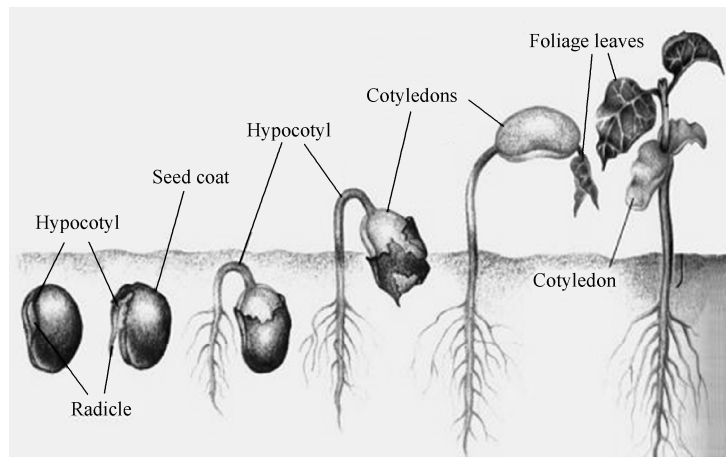


Figure 4.2 Seed germination

This is an example of a countably infinite sample space. In practice, there is some upper limit to the number of weeks anyone can live, but since this upper limit is unknown, we include all nonnegative integers in the sample space.

Example 4.7 Choosing a point from the interval $(0, 1)$. The sample space is $S = (0, 1)$. $E = (1/3, 1/2)$ is an event.

Example 4.8 Measure the lifetime of a lightbulb. The sample space is $S = [0, \infty)$. $E = [90, \infty)$ is an event.

That is

$$S = \mathbb{R}^+ \equiv (0, \infty)$$

This is an example of an uncountably infinite sample space.

4.1.3 Events

A *subset* of the sample space (a collection of possible outcomes) is known as an *event*. Events may be classified into four types:

- the *null event* is the empty subset of the sample space;
- an *atomic event* is a subset consisting of a single element of the sample space;
- a *compound event* is a subset consisting of more than one element of the sample space;
- the *sample space* itself is also an event.

Example 4.9

Consider all nonnegative integer points in number line (number of siblings),

$$S = \{0, 1, 2, \dots\}$$

and the event *at most two siblings*,

$$E = \{0, 1, 2\}.$$

Now consider the event

$$F = \{1, 2, 3, \dots\}.$$

Here, F is the event at least one sibling.

The *union* of two events E and F is the event that at least one of E and F occurs. The union of the events can be obtained by forming the union of the sets, see Figure 4.3. Thus, if G is the union of E and F , then we write

$$\begin{aligned} G &= E \cup F \\ &= \{0, 1, 2\} \cup \{1, 2, 3, \dots\} \\ &= \{0, 1, 2, \dots\} \\ &= S \end{aligned}$$

So the union of E and F is the whole sample space. That is, the events E and F together cover all possible outcomes of the experiment—at least one of E or F must occur.

The *intersection* of two events E and F is the event that both E and F occur. The intersection of two events can be obtained by forming the intersection of the sets, see Figure 4.3 and Figure 4.5. Thus, if H is the intersection of E and F , then

$$\begin{aligned} H &= E \cap F \\ &= \{0, 1, 2\} \cap \{1, 2, 3, \dots\} \\ &= \{1, 2\} \end{aligned}$$

So the intersection of E and F is the event *one or two siblings*.

The *complement* of an event, A , denoted A^c or \bar{A} , is the event that A does *not* occur, and hence consists of all those elements of the sample space that are not in A , see Figure 4.4. Thus if $E = \{0, 1, 2\}$ and $F = \{1, 2, \dots\}$,

$$E^c = \{3, 4, 5, \dots\}$$

and

$$F^c = \{0\}.$$

Two events A and B are *disjoint* or *mutually exclusive* if they cannot both occur. That is, their intersection is empty

$$A \cap B = \emptyset.$$

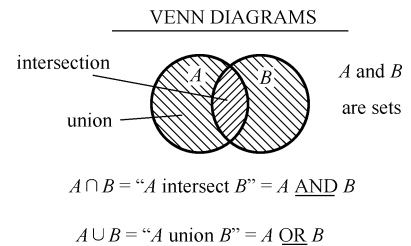


Figure 4.3 Union of two events and Intersection of two events

Note that for any event A , the events A and A^c are disjoint, and their union is the whole of the sample space:

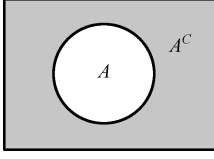


Figure 4.4 the events A and A^c

$$A \cap A^c = \emptyset \quad \text{and} \quad A \cup A^c = S.$$

The event A is *true* if the outcome of the experiment, s , is contained in the event A ; that is, if $s \in A$. We say that the event A *implies* the event B , and write $A \Rightarrow B$, if the truth of B automatically follows from the truth of A . If A is a subset of B , then occurrence of A necessarily implies occurrence of the event B . That is

$$(A \subseteq B) \iff (A \cap B = A) \iff (A \Rightarrow B).$$

We can see already that to understand events, we must understand a little set theory.

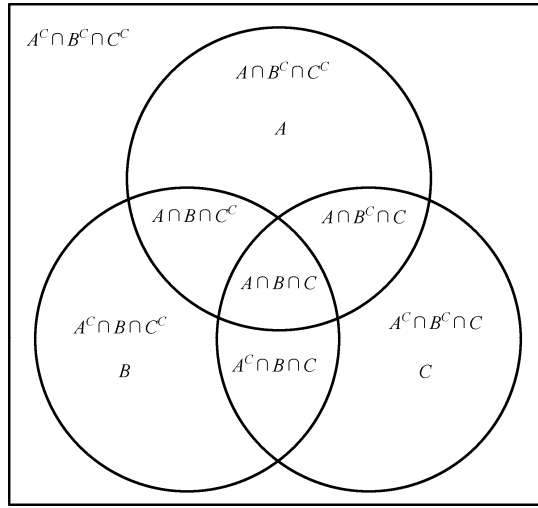


Figure 4.5 Intersection of three events

4.1.4 Set Theory

We already know about sets, complements of sets, and the union and intersection of two sets. In order to progress further we need to know the basic rules of set theory.

(I) Commutative laws:

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

(II) Associative laws:

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

(III) Distributive laws:

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

(IV) DeMorgan's laws:

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup B^c$$

(V) Disjoint union:

$$A \cup B = (A \cap B^c) \cup (A^c \cap B) \cup (A \cap B)$$

and $A \cap B^c$, $A^c \cap B$ and $A \cap B$ are disjoint.

Venn diagrams can be useful for thinking about manipulating sets, but formal proofs of set-theoretic relationships should only rely on use of the above laws.

New Words and Expressions

summarize ['sʌməraɪz] vt. 总结, 概述

germinate ['dʒɜːmɪneɪt] vt. 使发芽; 使发育; 使发展

monitor ['mɒnɪtə(r)] vt. 监控, 监听; 搜集, 记录; 测定

sibling ['sɪblɪŋ] n. 兄弟, 姐妹; [生]同科, 同属

Technical Terms

seed germination 种子发芽

countably infinite 可数无穷的, 可数无限的, 可数无穷多个

uncountably infinite 不可数无穷的, 不可数无限大的, 不可数无穷多个

upper limit 上限, 上限值

null event 零事件, 空事件

atomic event 原子事件

compound event 复合事件

mutually exclusive 互斥的; 不相容的; 互不相交的

Venn diagram 文氏图, 文恩图解

Commutative laws 交换律

Associative laws 结合律

Distributive laws 分配率

DeMorgan's laws 德摩根律

Disjoint union 不相交并集

4.2 Probability Axioms and Simple Counting Problems

4.2.1 Probability Axioms and Simple Properties

Probability is the language we use to model uncertainty. The data and examples we looked at many examples in this unit were the *outcomes* of *scientific experiments*. However, those outcomes could have been different—many different kinds of *uncertainty* and *randomness* were part of the mechanism which led to the actual data we saw. If we are to develop a proper understanding of such experimental results, we need to be able to understand the randomness underlying them.

With random phenomena, we can't predict the individual outcomes, but we hope to understand characteristics of their long-run behavior.

Now that we have a good mathematical framework for understanding events in terms of sets, we need a corresponding framework for understanding probabilities of events in terms of sets.

The *real valued function* $P(\cdot)$ is a *probability measure* if it acts on subsets of S and obeys the following Kolmogorov axioms:

(I) $P(S) = 1$;

(II) If $A \subseteq S$ then $P(A) \geq 0$;

(III) If A and B are *disjoint* ($A \cap B = \emptyset$) then

$$P(A \cup B) = P(A) + P(B).$$

Repeated use of Axiom III gives the more general result that if A_1, A_2, \dots, A_n are mutually disjoint, then

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i).$$

Indeed, we will assume further that the above result holds even if we have a *countably* infinite collection of disjoint events ($n = \infty$).

These axioms seem to fit well with our intuitive understanding of probability, but there are a few additional comments worth making.

1. Axiom I says that one of the possible outcomes must occur. A probability of 1 is assigned to the event “something occurs”. This fits in exactly with our definition of sample space. Note however, that the implication does not go the other way! When dealing with infinite sample spaces, there are often events of probability one which are not the sample space and events of probability zero which are not the empty set.

2. Axiom II simply states that we wish to work only with positive probabilities, because in some sense, probability measures the *size* of the set (event).

3. Axiom III says that probabilities “add up”—if we want to know the probability of *at most one sibling*, then this is the sum of the probabilities of *zero siblings* and *one sibling*. Allowing this result to hold for countably infinite unions is slightly controversial, but it makes the mathematics much easier, so we will assume it throughout!

These axioms are all we need to develop a theory of probability, but there are a collection of commonly used properties which follow directly from these axioms, and which we make extensive use of when carrying out probability calculations.

Property A: $P(A^c) = 1 - P(A)$.

Property B: $P(\emptyset) = 0$.

Property C: If $A \subseteq B$, then $P(A) \leq P(B)$.

Property D: (Addition Law) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

4.2.2 Interpretations of Probability

Somehow, we all have an intuitive feel for the notion of probability, and the axioms seem to capture its essence in a mathematical form. However, for probability theory to be anything other than an interesting piece of abstract pure mathematics, it must have an interpretation that in some way connects it to reality. If you wish only to study probability as a mathematical theory, then there is no need to have an interpretation.

However, if you are to use probability theory as your foundation for a theory of statistical inference which makes probabilistic statements about the world around us, then there must be an interpretation of probability which makes some connection between the mathematical theory and reality.

Whilst there is (almost) unanimous agreement about the mathematics of probability, the axioms and their consequences, there is considerable disagreement about the interpretation of probability. The three most common interpretations are given below.

(1) Classical interpretation

The classical interpretation of probability is based on the assumption of underlying equally likely events. That is, for any events under consideration, there is always a sample space which can be considered where all atomic events are equally likely. If this sample space is given, then the probability axioms may be deduced from set-theoretic considerations.

This interpretation is fine when it is obvious how to partition the sample space into equally likely events, and is in fact entirely compatible with the other two interpretations to be described in that case. The problem with this interpretation is that for many situations it is not at all obvious what the partition into equally likely events is. For example, consider the probability that it rains in Newcastle tomorrow. This is clearly a reasonable event to consider, but it is not at all clear what sample space we should construct with equally likely outcomes. Consequently, the classical interpretation falls short of being a good interpretation for real-world problems. However, it provides a good starting point for a mathematical treatment of probability theory, and is the interpretation adopted by many mathematicians and theoreticians.

(2) Frequentist interpretation

An interpretation of probability widely adopted by statisticians is the relative frequency interpretation. This interpretation makes a much stronger connection with reality than the previous one, and fits in well with traditional statistical methodology. Here probability only has meaning for events from experiments which could in principle be repeated arbitrarily many times under essentially identical conditions. Here, the probability of an event is simply the “long-run proportion” of times that the event occurs under many repetitions of the experiment. It is reasonable to suppose that this proportion will settle down to some limiting value eventually, which is the probability of the event. In such a situation, it is possible to derive the axioms of probability from consideration of the long run frequencies of various events. The probability p , of an event E , is defined by

$$p = \lim_{n \rightarrow \infty} \frac{r}{n}$$

where r is the number of times E occurred in n repetitions of the experiment.

Unfortunately it is hard to make precise exactly why such a limiting frequency should exist. A bigger problem however, is that the interpretation only applies to outcomes of repeatable experiments, and there are many “one-off” events, such as “rain in Newcastle tomorrow”, that we would like to be able to attach probabilities to.

(3) Subjective interpretation

This final common interpretation of probability is somewhat controversial, but does not suffer from the problems that the other interpretations do. It suggests that the association of probabilities to events is a personal (subjective) process, relating to your *degree of belief* in the likelihood of the event occurring. It is controversial because it accepts that *different* people will assign *different* probabilities to the *same event*. Whilst in some sense it gives up on an objective notion of probability, it is in no sense arbitrary. It can be defined in a precise way, from which the axioms of probability may be derived as requirements of self-consistency.

A simple way to define *your* subjective probability that some event E will occur is as follows. Your probability is the number p such that you consider £ p to be a *fair price* for a gamble which will pay you £1 if E occurs and nothing otherwise.

So, if you consider 40p to be a fair price for a gamble which pays you £1 if it rains in Newcastle tomorrow, then 0.4 is your subjective probability for the event. The subjective interpretation is sometimes known as the *degree of belief interpretation*, and is the interpretation of probability underlying the theory of *Bayesian Statistics*—a powerful theory of statistical inference named after Thomas Bayes, the 18th Century Presbyterian Minister who first proposed it. Consequently, this interpretation of probability is sometimes also known as the *Bayesian interpretation*.

Summary

Whilst the interpretation of probability is philosophically very important, all interpretations lead to the same set of axioms, from which the rest of probability theory is deduced. Consequently, for this section, it will be sufficient to adopt a fairly classical approach, taking the axioms as given, and investigating their consequences independently of the precise interpretation adopted.

4.2.3 Classical Probability

Classical probability theory is concerned with carrying out probability calculations based on *equally likely outcomes*. That is, it is assumed that the sample space has been constructed in such a way that every subset of the sample space consisting of a single element has the same probability, see Figure 4.6. If the sample space contains n possible outcomes ($\#S = n$), we must have for all $s \in S$,

$$P(\{s\}) = \frac{1}{n}$$

and hence for all $E \subseteq S$

$$P(E) = \frac{\#E}{n}.$$

More informally, we have

$$P(E) = \frac{\text{number of ways } E \text{ can occur}}{\text{total number of outcomes}}$$

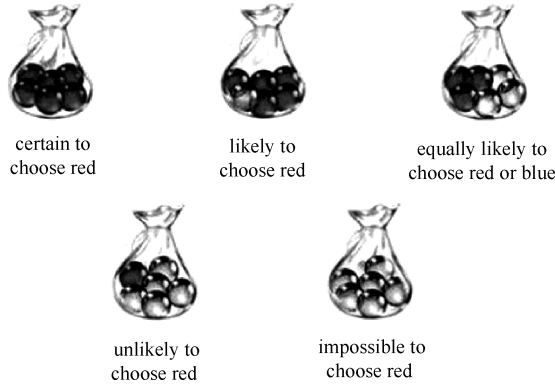


Figure 4.6 Many ways of possible chooses

Example 4.10

Suppose that a fair coin is thrown twice, and the results recorded. The sample space is

$$S = \{HH, HT, TH, TT\}.$$

Let us assume that each outcome is equally likely—that is, each outcome has a probability of $1/4$. Let A denote the event *head on the first toss*, and B denote the event *head on the second toss*. In terms of sets

$$A = \{HH, HT\}, B = \{HH, TH\}.$$

So

$$P(A) = \frac{\#A}{n} = \frac{2}{4} = \frac{1}{2}$$

and similarly $P(B) = 1/2$. If we are interested in the event $C = A \cup B$ we can work out its probability using from the set definition as

$$P(C) = \frac{\#C}{4} = \frac{\#(A \cup B)}{4} = \frac{\#\{HH, HT, TH\}}{4} = \frac{3}{4}$$

or by using the addition formula

$$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - P(A \cap B).$$

Now $A \cap B = \{HH\}$, which has probability $1/4$, so

$$P(C) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}.$$

In this simple example, it seems easier to work directly with the definition. However, in more complex problems, it is usually much easier to work out how many elements there are in an intersection than in a union, making the addition law very useful.

4.2.4 The Multiplication Principle

In the above example we saw that there were two distinct experiments—*first throw* and

second throw. There were two equally likely outcomes for the first throw and two equally likely outcomes for the second throw. This leads to a combined experiment with $2 \times 2 = 4$ possible outcomes. This is an example of the *multiplication principle*.

Multiplication principle

If there are p experiments and the first has n_1 equally likely outcomes, the second has n_2 equally likely outcomes, and so on until the p th experiment has n_p equally likely outcomes, then there are

$$n_1 \cdot n_2 \cdots n_p = \prod_{i=1}^p n_i$$

equally likely possible outcomes for the p experiments.

Example 4.11

A class of school children consists of 14 boys and 17 girls. The teacher wishes to pick one boy and one girl to star in the school play. By the multiplication principle, she can do this in $14 \times 17 = 238$ different ways.

Example 4.12

A die is thrown twice and the number on each throw is recorded. There are clearly 6 possible outcomes for the first throw and 6 for the second throw. By the multiplication principle, there are 36 possible outcomes for the two throws. If D is the event *a double-six*, then since there is only one possible outcome of the two throws which leads to a double-six, we must have $P(D) = 1/36$.

Now let E be the event *six on the first throw* and F be the event *six on the second throw*. We know that $P(E) = P(F) = 1/6$. If we are interested in the event G , *at least one six*, then $G = E \cup F$, and using the addition law we have

$$\begin{aligned} P(G) &= P(E \cup F) \\ &= P(E) + P(F) - P(E \cap F) \\ &= \frac{1}{6} + \frac{1}{6} - P(D) \\ &= \frac{1}{6} + \frac{1}{6} - \frac{1}{36} \\ &= \frac{11}{36}. \end{aligned}$$

This is much easier than trying to count how many of the 36 possible outcomes correspond to G .

New Words and Expressions

obey [ə'beɪ] vt. & vi. 服从, 听从

double-six 双六

multiplication principle 乘法原理

4.3.2 Permutations

Suppose that we have a collection of n objects, $C = \{c_1, c_2, \dots, c_n\}$. We want to make r selections from C . How many possible *ordered* selections can we make?

If we are sampling *with replacement*, then we have r experiments, and each has n possible (equally likely) outcomes, and so by the multiplication principle, there are

$$n \cdot n \cdot \dots \cdot n = n^r$$

ways of doing this.

If we are sampling *without replacement*, then we have r experiments. The first experiment has n possible outcomes. The second experiment only has $n - 1$ possible outcomes, as one object has already been selected. The third experiment has $n - 2$ outcomes and so on until the r th experiment, which has $n - r + 1$ possible outcomes. By the multiplication principle, the number of possible selections is

$$n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot (n-r+1) = \frac{n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1}{(n-r) \cdot (n-r-1) \cdot \dots \cdot 3 \cdot 2 \cdot 1} = \frac{n!}{(n-r)!}$$

This is a commonly encountered expression in combinatorics, and has its own notation. The number of ordered ways of selecting r objects from n is denoted P_r^n ,

$$P_r^n = \frac{n!}{(n-r)!}.$$

We refer to P_r^n as the number of permutations of r out of n objects. If we are interested solely in the number of ways of arranging n objects, then this is clearly just, see Figure 4.7.

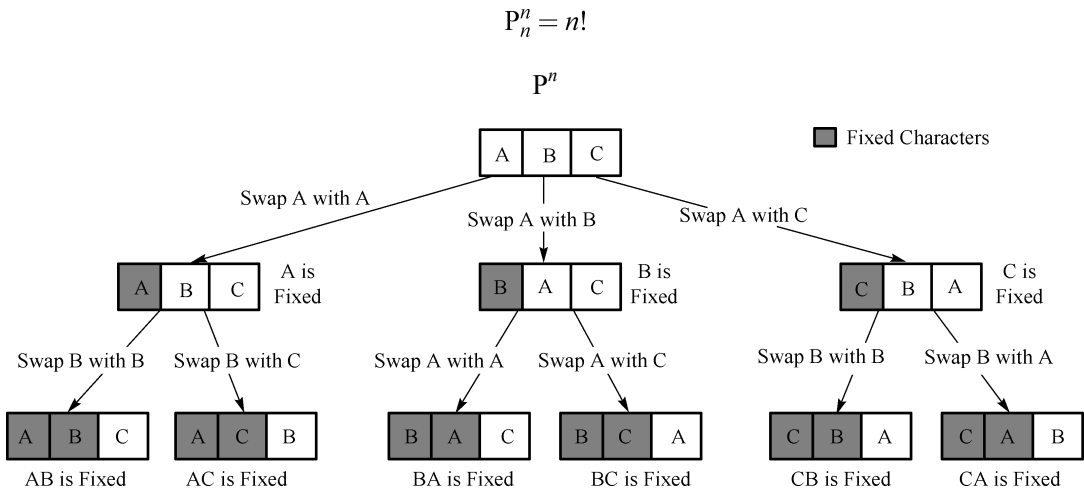


Figure 4.7 Recursion tree for permutations of string “ABC”

Example 4.13

A CD has 12 tracks on it, and these are to be played in random order. There are $12!$ ways of selecting them. There is only one such ordering corresponding to the ordering on the box, so the

probability of the tracks being played in the order on the box is $1/12!$! As we will see later, this is considerably smaller than the probability of winning the National Lottery!

Suppose that you have time to listen to only 5 tracks before you go out. There are

$$P_5^{12} = \frac{12!}{7!} = 12 \times 11 \times 10 \times 9 \times 8 = 95,040$$

ways they could be played. Again, only one of these will correspond to the first 5 tracks on the box (in the correct order), so the probability that the 5 played will be the first 5 on the box is $1/95040$.

Example 4.14 Two students share a birthday

In a computer practical session containing 40 students, what is the probability that at least two students share a birthday?

First, let's make some simplifying assumptions. We will assume that there are 365 days in a year and that each day is equally likely to be a birthday.

Call the event we are interested in A . We will first calculate the probability of A^c , the probability that *no two* people have the same birthday, and calculate the probability we want using $P(A) = 1 - P(A^c)$. The number of ways 40 birthdays could occur is like sampling 40 objects from 365 *with* replacement, which is just 365^{40} . The number of ways we can have 40 *distinct* birthdays is like sampling 40 objects from 365 *without* replacement, P_{40}^{365} . So, the probability of all birthdays being distinct is

$$P(A^c) = \frac{P_{40}^{365}}{365^{40}} = \frac{365!}{325!365^{40}} \approx 0.1$$

and so

$$P(A) = 1 - P(A^c) \approx 0.9.$$

That is, there is a probability of 0.9 that we have a match. In fact, the fact that birthdays are *not* distributed uniformly over the year makes the probability of a match even higher!

Suppose that you are one of a group of 40 students. What is the probability of B , where B is the event that at least one other person in the group has the same birthday as you?

Again, we will work out $P(B^c)$ first, the probability that no-one has your birthday. Now, there are 365^{39} ways that the birthdays of the other people can occur, and we allow each of them to have any birthday other than yours, so there are 364^{39} ways for this to occur. Hence we have

$$P(B^c) = \frac{364^{39}}{365^{39}} \approx 0.9$$

and so

$$P(B) = 1 - \frac{364^{39}}{365^{39}} \approx 0.1.$$

Here the probabilities are reversed—there is only a 10% chance that someone has the same birthday as you. Most people find this much more intuitively reasonable. So, how big a group of

people would you need in order to have a better than even chance of someone having the same birthday as you? The general formula for the probability of a match with n people is

$$P(B) = 1 - \frac{364^{n-1}}{365^{n-1}} = 1 - \left(\frac{364}{365}\right)^{n-1},$$

and as long as you enter it into your calculator the way it is written on the right, it will be fine. We find that a group of size 254 is needed for the probability to be greater than 0.5, and that a group of 800 or more is needed before you can be really confident that someone will have the same birthday as you. For a group of size 150 (the size of the lectures), the probability of a match is about 1/3.

This problem illustrates quite nicely the subtlety of probability questions, the need to define precisely the events you are interested in, and the fact that some probability questions have counter-intuitive answers.

4.3.3 Combinations

We now have a way of counting permutations, but often when selecting objects, all that matters is *which* objects were selected, not the order in which they were selected. Suppose that we have a collection of objects, $C = \{c_1, \dots, c_n\}$ and that we wish to make r selections from this list of objects, *without replacement*, where the order does not matter. An unordered selection such as this is referred to as a *combination*. How many ways can this be done? Notice that this is equivalent to asking how many different subsets of C of size r there are.

From the multiplication principle, we know that the number of *ordered* samples must be the number of *unordered* samples, multiplied by the number of orderings of each sample. So, the number of unordered samples is the number of ordered samples, divided by the number of orderings of each sample. That is, the number of unordered samples is

$$\begin{aligned} \frac{\text{number of ordered samples of size } r}{\text{number of orderings of samples of size } r} &= \frac{P_r^n}{P_r^r} \\ &= \frac{P_r^n}{r!} \\ &= \frac{n!}{r!(n-r)!} \end{aligned}$$

Again, this is a very commonly found expression in combinatorics, so it has its own notation. In fact, there are two commonly used expressions for this quantity:

$$C_r^n = \binom{n}{r} = \frac{n!}{r!(n-r)!}.$$

These numbers are known as the *binomial coefficients*. We will use the notation $\binom{n}{r}$ as this is slightly neater, and more commonly used. They can be found as the $(r+1)$ th number on the $(n+1)$ th row of Pascal's triangle:

				1				
				1		1		
			1	2		1		
		1	3	3		1		
	1	4	6	4		1		
1	5	10	10	5		1		
1	6	15	20	15		6		1
⋮			⋮					⋮

Notes: the binomial coefficients in general be written as $\binom{n}{k} = \frac{n!}{k!(n-k)!}$.

(I) This is read “ n choose k ”.

(II) The “!” means factorial.

(III) n is the total number of observations.

(IV) k is the number of observations you are interested in.

Example 4.15

Return to the CD with 12 tracks. You arrange for your CD player to play 5 tracks at random. How many different unordered selections of 5 tracks are there, and what is the probability that the 5 tracks played are your 5 favourite tracks (in any order)?

The number of ways of choosing 5 tracks from 12 is just $\binom{12}{5} = 792$. Since only one of these will correspond to your favourite five, the probability of getting your favourite five is $1/792 \approx 0.001$.

Example 4.16 National Lottery

What is the probability of winning exactly £10 on the National Lottery?

In the UK National Lottery, there are 49 numbered balls, and six of these are selected at random. A seventh ball is also selected, but this is only relevant if you get exactly five numbers correct. The player selects six numbers before the draw is made, and after the draw, counts how many numbers are in common with those drawn. If the player has selected exactly three of the balls drawn, then the player wins £10. The order the balls are drawn in is irrelevant.

We are interested in the probability that exactly 3 of the 6 numbers we select are drawn. First we need to count the number of possible draws (the number of different sets of 6 numbers), and then how many of those draws correspond to getting exactly three numbers correct. The number of possible draws is the number of ways of choosing 6 objects from 49. This is

$$\binom{49}{6} = 13,983,816.$$

The number of drawings corresponding to getting exactly three right is calculated as follows. Regard the six numbers you have chosen as your “good” numbers. Then of the 49 balls to be drawn from, 6 correspond to your “good” numbers, and 43 correspond to your “bad” numbers. We want to know how many ways there are of selecting 3 “good” numbers and 3 “bad” numbers. By the multiplication principle, this is the number of ways of choosing 3 from 6, multiplied by the number of ways of choosing 3 from 43. That is, there are

$$\binom{6}{3} \binom{43}{3} = 246,820$$

ways of choosing exactly 3 “good” numbers. So, the probability of getting exactly 3 numbers, and winning £10 is

$$\frac{\binom{6}{3} \binom{43}{3}}{\binom{49}{6}} \approx 0.0177 \approx \frac{1}{57}.$$



Figure 4.8 UK National Lottery

4.3.4 The Difference Between Permutations and Combinations

It is very important to make the distinction between permutations and combinations.

(i) In permutations, order matters and in combinations order does not matter. The important information can be summarized by:

Table 4.1 The Difference Between Permutations and Combinations

	Order	With repetition	Without repetition
Permutations	matters	$C_r^n = \frac{(n+r-1)!}{r!(n-1)!}$	$C_r^n = \frac{n!}{r!(n-1)!}$
Combinations	does not matter	$P_r^n = n^r$	$P_r^n = \frac{n!}{(n-1)!}$

Where:

A combination: is unordered list of items

A permutation: is an ordered list of items

Repetition: refer to whether an item on the list can repeat or not

n is the number of possible items that can be selected

r is the number of items that can be selected

(ii) Permutations and Combinations

Table 4.2 Permutations and Combinations

Permutations	Combinations
- Order of people/objects matters - “Hod words” arrangements/orders/rank (1 st 2 nd 3 rd place)	- Order of people/objects <u>does not</u> matter
Examples - any type of race (w/rank) - playlist/order on a shelf - student’ gov t (P, VP, S, T) - quanstion uses the “Hod Words”	Examples - people sitting in a car/vahicle - handshakes or high fives with classmates/teammates - any type of race (no rank, top 10 get t – shirts) - student senators

Example 4.17

A company has to select 3 officers from a pool of 6 candidates. How many different ways can this be done if (a) The officers are distinct? (b) The officers are not distinct?

It is very important whether or not these officers are distinct.

(a) If the officers are distinct, we are picking a triple (s_1, s_2, s_3) with each s_i being a candidate, and order matters. This means we are finding a 3-permutation from a set of 6 elements. So there are:

$$P_3^6 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = 6 \times 5 \times 4 = 120$$

distinct ways to pick these officers.

(b) If the officers are not distinct, the triples (s_1, s_2, s_3) , (s_1, s_3, s_2) , (s_3, s_2, s_1) , etc. are the same since the positions are the same. So, we are finding a 3-combinations from a set of 6 elements. So there are:

$$C_3^6 = \frac{6!}{(6-3)!3!} = \frac{6!}{3!3!} = \frac{6 \times 5 \times 4}{3 \times 2 \times 1} = 20$$

ways to pick these officers.

New Words and Expressions

duplicate ['dju:plɪkət] v. 重复; 复制; 复印

urn [ɜ:n] n. 大茶壶; 瓮; 缸

vase [vɑ:z] n. 花瓶; (装饰用的) 瓶

permutation [ˌpɜ:mju'teɪʃn] n. 序列, 排列, 排列中的任一组数字或文字

track [træk] n. 轨道, 音轨; 小路, 小道; 方针, 路线

session ['seʃn] n. 开会, 会议; (法庭的) 开庭; 会期, 学期

subtlety ['sʌtlɪti] n. 精妙, 巧妙; 敏锐, 敏感; 细微的差别等

neat [ni:t] adj. 整洁的, 干净的; 灵巧的; 匀整的

favourite ['feɪvərɪt] n. 特别喜爱的人(或物) adj. 特别受喜爱的

handshake ['hændʃeɪk] n. 握手, 复数 handshakes

playlist ['pleɪlɪst] n. 播放列表

Technical Terms

repeated sampling 重复抽样

sampling with replacement 放回抽样

sampling without replacement 不放回抽样

Pascal' triangle 帕斯卡三角; 帕斯卡三角形; 杨辉三角

Notes

the National Lottery (UK National Lottery) 英国国家彩票; 英国国家乐透

combinatorics 组合数学, 组合学

high five [hai farv] *n.* (指庆祝成功、表示致意等) 举手击掌, 击掌代表的意思随不同的语境而有所变化, 不过基本都是问候、祝贺或者庆祝的意思。这是一种美国文化手势, 一般代表了“庆祝成功的击掌”, 有时也写成“Give me five”。这个手势用于两人之间, 动作是两人各高举一只手, 并向对方的手拍击。

4.4 Conditional Probability and the Multiplication Rule

4.4.1 Conditional Probability

We now have a way of understanding the probabilities of events, but so far we have no way of *modifying* those probabilities when certain events occur. For this, we need an extra axiom which can be justified under any of the interpretations of probability.

The axiom defines the *conditional probability of A given B*, written $P(A|B)$ as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{for } P(B) > 0.$$

Note that we can only condition on events with positive probability.

Under the classical interpretation of probability, we can see that if we are told that B has occurred, then all outcomes in B are equally likely, and all outcomes not in B have zero probability—so B is the new sample space. The number of ways that A can occur is now just the number of ways $A \cap B$ can occur, and these are all equally likely. Consequently we have

$$P(A|B) = \frac{\#(A \cap B)}{\#B} = \frac{\#(A \cap B) / \#S}{\#B / \#S} = \frac{P(A \cap B)}{P(B)}.$$

Because conditional probabilities really just correspond to a new probability measure defined on a smaller sample space, they obey all of the properties of “ordinary” probabilities. For example, we have

$$P(B|B) = 1$$

$$P(\emptyset|B) = 0$$

$$P(A \cup C|B) = P(A|B) + P(C|B), \quad \text{for } A \cap C = \emptyset$$

and so on.

The definition of conditional probability simplifies when one event is a special case of the other. If $A \subseteq B$, then $A \cap B = A$ so

$$P(A|B) = \frac{P(A)}{P(B)}, \quad \text{for } A \subseteq B.$$

Example 4.18

A die is rolled and the number showing recorded, see Figure 4.9. Given that the number rolled was even, what is the probability that it was a six?

Let E denote the event “even” and F denote the event “a six”. Clearly $F \subseteq E$, so

$$P(F|E) = \frac{P(E)}{P(E)} = \frac{1/6}{1/2} = \frac{1}{3}.$$



Figure 4.9 UK National Lottery

4.4.2 The Multiplication Rule

The formula for conditional probability is useful when we want to calculate $P(A|B)$ from $P(A \cap B)$ and $P(B)$. However, more commonly we want to know $P(A \cap B)$ and we know $P(A|B)$ and $P(B)$. A simple rearrangement gives us the multiplication rule.

$$P(A \cap B) = P(B) \times P(A|B)$$

Example 4.19

Two cards are dealt from a deck of 52 cards. What is the probability that they are both Aces?

We now have three different ways of computing this probability. First, let's use conditional probability. Let A_1 be the event “first card an Ace” and A_2 be the event “second card an Ace”. $P(A_2|A_1)$ is the probability of a second Ace. Given that the first card has been drawn and was an Ace, there are 51 cards left, 3 of which are Aces, so $P(A_2|A_1) = 3/51$. So,

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_1) \cdot P(A_2|A_1) \\ &= \frac{4}{52} \cdot \frac{3}{51} \\ &= \frac{1}{221}. \end{aligned}$$

Now let's compute it by counting ordered possibilities. There are P_2^{52} ways of choosing 2 cards from 52, and P_2^4 of those ways correspond to choosing 2 Aces from 4, so

$$P(A_1 \cap A_2) = \frac{P_2^4}{P_2^{52}} = \frac{12}{2652} = \frac{1}{221}.$$

Now let's compute it by counting unordered possibilities. There are $\binom{52}{2}$ ways of choosing 2 cards from 52, and $\binom{4}{2}$ of those ways correspond to choosing 2 Aces from 4, so

$$P(A_1 \cap A_2) = \frac{\binom{4}{2}}{\binom{52}{2}} = \frac{6}{1326} = \frac{1}{221}.$$

If possible, you should always try and calculate probabilities more than one way (as it is very easy to go wrong!). However, for counting problems where the order doesn't matter, counting the unordered possibilities using combinations will often be the only reasonable way, and for problems which don't correspond to a sampling experiment, using conditional probability will often be the only reasonable way.

The multiplication rule generalizes to more than two events. For example, for three events we have

$$P(A_1 \cap A_2 \cap A_3) = P(A_1)P(A_2 | A_1)P(A_3 | A_1 \cap A_2).$$

New Words and Expressions

modify ['mɒdɪfaɪ] *vi.* 被修饰; 修改 *vt.* 改变; 减轻, 减缓

roll [rəʊl] *vt.* 辗; (使) 原地转圈; 滚动

even ['i:vən] *adj.* 偶数的; 公平的; 平坦的; 平均的

desk [desk] *n.* 书桌, 办公桌; 服务台; 部门 *adj.* 书桌的, 书桌上用的

card [kɑ:d] *n.* 纸牌; 卡片; 明信片; 信用卡

Ace [eɪs] *n.* A 纸牌 (亦称“爱司”); 擅长……的人; 精于……的人

generalize ['dʒenərəlaɪz] *vt.* 推广, 普及

Technical Terms

conditional probability 条件概率

multiplication rule 乘法规则

sampling experiment 抽样试验

4.5 Independent Events, Partitions and Bayes Theorem

4.5.1 Independence

Recall the multiplication rule

$$P(A \cap B) = P(B)P(A | B).$$

For some events A and B , knowing that B has occurred will not alter the probability of A , so that $P(A|B) = P(A)$. When this is so, the multiplication rule becomes

$$P(A \cap B) = P(A)P(B),$$

and the events A and B are said to be *independent events*. Independence is a very important concept in probability theory, and is used a lot to build up complex events from simple ones. Do not confuse the independence of A and B with the exclusivity of A and B —they are entirely different concepts. If A and B both have positive probability, then they cannot be both independent and exclusive (exercise).

When it is clear that the occurrence of B can have no influence on A , we will *assume* independence in order to calculate $P(A \cap B)$. However, if we can calculate $P(A \cap B)$ directly, we can check the independence of A and B by seeing if it is true that

$$P(A \cap B) = P(A)P(B).$$

We can generalize independence to collections of events as follows.

The set of events $A = \{A_1, A_2, \dots, A_n\}$ are *mutually independent events* if for any subset, $B \subseteq A$, $B = \{B_1, B_2, \dots, B_r\}$, $r \leq n$ we have

$$P(B_1 \cap \dots \cap B_r) = P(B_1) \cdots P(B_r).$$

Note that mutual independence is much stronger than *pair-wise* independence, where we only require independence of subsets of size 2. That is, pair-wise independence *does not* imply mutual independence.

Example 4.20

A playing card is drawn from a pack. Let A be the event “an Ace is drawn” and let C be the event “a Club is drawn”. Are the events A and C exclusive? Are they independent?

A and C are clearly not exclusive, since they can both happen—when the Ace of Clubs is drawn. Indeed, since this is the only way it can happen, we know that $P(A \cap C) = 1/52$. We also know that $P(A) = 1/13$ and that $P(C) = 1/4$. Now since

$$\begin{aligned} P(A)P(C) &= \frac{1}{13} \cdot \frac{1}{4} \\ &= \frac{1}{52} \\ &= P(A \cap C) \end{aligned}$$

we know that A and C are independent. Of course, this is intuitively obvious—you are no more or less likely to think you have an Ace if someone tells you that you have a Club.

4.5.2 Partitions

A *partition* of a sample space is simply the decomposition of the sample space into a collection of mutually *exclusive* events with positive probability, see Figure 4.10. That is, $\{B_1, \dots, B_n\}$ form a *partition* of S if

- (i) $S = B_1 \cup B_2 \cup \dots \cup B_n = \bigcup_{i=1}^n B_i$,
- (ii) $B_i \cap B_j = \emptyset$, $\forall i \neq j$,
- (iii) $P(B_i) > 0$, $\forall i$.

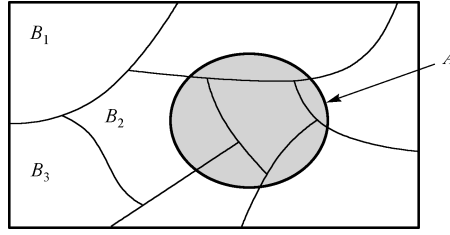


Figure 4.10 A partition $\{B_1, \dots, B_n\}$ of S

Example 4.21

A card is randomly drawn from the pack. The events $\{C, D, H, S\}$ (Club, Diamond, Heart, Spade) form a partition of the sample space, since one and only one will occur, and all can occur.

4.5.3 Law of Total Probability

Suppose that we have a partition $\{B_1, \dots, B_n\}$ of a sample space, S . Suppose further that we have an event A . Then A can be written as the disjoint union

$$A = (A \cap B_1) \cup \dots \cup (A \cap B_n),$$

and so the probability of A is given by

$$\begin{aligned} P(A) &= P((A \cap B_1) \cup \dots \cup (A \cap B_n)) \\ &= P(A \cap B_1) + \dots + P(A \cap B_n), \quad \text{by Axiom III} \\ &= P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n), \quad \text{by the multiplication rule} \\ &= \sum_{i=1}^n P(A|B_i)P(B_i). \end{aligned}$$

Law of Total Probability:

If B_1, B_2, B_3, \dots is a partition of the sample space S , then for any event A We have

$$P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$$

Example 4.21 “Craps”

Craps is a game played with a pair of dice. A player plays against a banker. The player throws the dice and notes the sum.

- If the sum is 7 or 11, the player wins, and the game ends (a *natural*).
- If the sum is 2, 3 or 12, the player loses and the game ends (a *crap*).
- If the sum is anything else, the sum is called the players *point*, and the player keeps throwing the dice until his sum is 7, in which case he loses, or he throws his *point* again, in which case he wins.

What is the probability that the player wins?

4.5.4 Bayes Theorem

From the multiplication rule, we know that

$$P(A \cap B) = P(B)P(A|B)$$

and that

$$P(A \cap B) = P(A)P(B|A),$$

so clearly

$$P(B)P(A|B) = P(A)P(B|A),$$

and so

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

This is known as *Bayes Theorem*, and is a very important result in probability, as it tells us how to “turn conditional probabilities around”—that is, it tells us how to work out $P(A|B)$ from $P(B|A)$, and this is often very useful.

Example 4.23

A clinic offers you a free test for a very rare, but hideous disease. The test they offer is very reliable. If you have the disease it has a 98% chance of giving a positive result, and if you don’t have the disease, it has only a 1% chance of giving a positive result. You decide to take the test, and find that you test positive—what is the probability that you have the disease?

Let P be the event “test positive” and D be the event “you have the disease”. We know that

$$P(P|D) = 0.98 \text{ and then } P(P|D^c) = 0.01.$$

We want to know $P(D|P)$, so we use Bayes’ Theorem.

$$\begin{aligned} P(D|P) &= \frac{P(P|D)P(D)}{P(P)} \\ &= \frac{P(P|D)P(D)}{P(P|D)P(D) + P(P|D^c)P(D^c)} \quad (\text{using the theorem of total probability}) \\ &= \frac{0.98P(D)}{0.98P(D) + 0.01(1 - P(D))}. \end{aligned}$$

So we see that the probability you have the disease given the test result depends on the probability that you had the disease in the first place. This is a rare disease, affecting only one in ten thousand people, so that $P(D) = 0.0001$. Substituting this in gives

$$P(D|P) = \frac{0.98 \times 0.0001}{0.98 \times 0.0001 + 0.01 \times 0.9999} \approx 0.01.$$

So, your probability of having the disease has increased from 1 in 10,000 to 1 in 100, but still isn’t that much to get worried about! Note the *crucial* difference between $P(P|D)$ and $P(D|P)$.

4.5.5 Bayes Theorem for Partitions

Another important thing to notice about the above example is the use of the theorem of total

probability in order to expand the bottom line of Bayes Theorem. In fact, this is done so often that Bayes Theorem is often stated in this form.

Suppose that we have a partition $\{B_1, \dots, B_n\}$ of a sample space S . Suppose further that we have an event A , with $P(A) > 0$. Then, for each B_j , the probability of B_j given A is

$$\begin{aligned} P(B_j|A) &= \frac{P(A|B_j)}{P(A)} \\ &= \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \dots + P(A|B_n)P(B_n)} \\ &= \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^n P(A|B_i)P(B_i)}. \end{aligned}$$

In particular, if the partition is simply $\{B, B^c\}$, then this simplifies to

$$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}.$$

New Words and Expressions

independence [ˌɪndɪˈpendəns] *n.* 独立; 独立性; 独立心

exclusivity [ˌɛkskluːsɪvəti] *n.* 排他性; 专有权; 独特性

pair-wise [peərˈwaɪz] *adj.* 两个两个的; 两两的

pack [pæk] *n.* (纸牌的)一副; 一群; 包裹; 一组

Club [klʌb] *n.* (纸牌的)梅花; 俱乐部, 会所; 社团; 夜总会

Diamond [ˈdaɪəmənd] *n.* (纸牌的)方块; 钻石, 金刚石; 菱形

Heart [hɑ:t] *n.* (纸牌的)红桃; 心, 内心; 核心; 心脏

Spade [speɪd] *n.* (纸牌的)黑桃; 铁锹, 铲子

craps [kræps] [美]双骰子赌博; 花旗骰 (Craps), 别名: 派司拉 (Pass Line, 在澳门新葡京娱乐场所“中文”释本)

dice [daɪs] *n.* 骰子; 掷骰游戏

banker [ˈbæŋkə(r)] *n.* 庄家; 银行家; 银行经理

clinic [ˈklɪnɪk] *n.* 诊所, 门诊部

hideous [ˈhɪdiəs] *adj.* 令人惊骇的; 极其丑陋的, 可怕的; 丑恶的, 讨厌的

crucial [ˈkruːʃl] *adj.* 至关重要的, 决定性的; 关键性的; 严重的

Technical Terms

independent event 独立事件

mutually independent events 相互独立事件

pairwise independent events 两两独立事件；成对独立事件

mutually exclusive events 互斥事件，互不相容事件，不相交事件

Notes

1. Banker and Player: 赌场里的庄家(Banker)。这个词是赌场专用语，指开局设赌者，通常只有赌场才具备这一资格，与闲家(Player)对赌。

2. What does craps mean? A gambling game played with two dice; a first throw of 7 or 11 wins and a first throw of 2, 3, or 12 loses and a first throw of any other number must be repeated to win before a 7 is thrown, which loses the bet and the dice.

Passage 1. Probability and Odds

Probabilities can be and are expressed in many ways; we see and hear many of them in the news nearly every day. Odds (发生比, 优势比或几率) are a way of expressing probabilities by expressing the number of ways an event can happen compared with the number of ways it cannot happen. The statement “It is four times more likely to rain tomorrow than not rain” is a probability statement and is expressed as odds: “The odds are 4 to 1 in favor of rain tomorrow” (also written 4:1).

Odds are another way of expressing probability. For some event A that occurs with probability p , the “odds of A ” are the ratio of (the probability of) A happening to A not-happening, so the odds of event A are $p/(1-p)$.

If there is a p probability of something happening, then the odds can be considered the number of successes you expect to get for every failure on average. High odds correspond to high probabilities, low odds to low probabilities.

To calculate the odds:

$$O = \frac{\text{proportion of successes}}{\text{proportion of failures}} = \frac{p}{1-p}$$

To go from odds back to probabilities:

$$p = \frac{O}{1+O}$$

While probabilities are bound to $[0, 1]$, odds are bound to $[0, \infty)$. Let's take the titanic example (泰坦尼克号例子). What was the overall odds of survival on the Titanic. 38% of passengers survived the Titanic. Therefore:

$$\text{Odds of survival} = \frac{0.38}{1-0.38} = \frac{0.38}{0.62} = 0.61$$

How can we say this in English: “For every one death on the Titanic, there was on average 0.61 survivors.”

Passage 2. The Relationship between Odds and Probability

If the odds in favor of event A are a to b (or $a:b$), then

(i) The odds against event A are b to a (or $b:a$).

(ii) The probability of event A is $p(A) = \frac{a}{a+b}$.

(iii) The probability that event A will not occur is $p(\text{not } A) = \frac{b}{a+b}$.

To illustrate this relationship, consider the statement “The odds favoring rain tomorrow are 4: 1.” Using the preceding notation, $a = 4$ and $b = 1$.

Therefore, the probability of rain tomorrow is $\frac{4}{4+1}$, or $4/5 = 0.8$. the odds against rain tomorrow are 1 to 4 (or 1:4), and the probability that there is no rain tomorrow is $\frac{1}{4+1}$, or $1/5=0.2$.

Passage 3. How the Odds Change across the Range of the Probability

Let’s see how the odds change across the range of the probability.

p	0	0.01	0.05	0.10	0.3333	0.40	0.50	0.6	0.6667	0.75	0.9	0.95	0.999
O	0	0.010	0.052	0.111	0.5	0.6667	1	1.5	2	3	9	19	999

Odds ratio: the ratio of two odds. If one odds is twice as big as the other, then we say the odds ratio is 2. If it is half again as big, the odds ratio would be 1.5.

$$OR = O_1 / O_2 = \frac{p_1}{1-p_1} \bigg/ \frac{p_2}{1-p_2}$$

Let’s go back to the Titanic example. Let’s look at the odds ratio between men and women on the Titanic. 19% of men survived while 73% of women survived. Let’s look at this as a bar graph, see Figure 4.11.

Now let’s calculate the odds of survival for each group,

$$\text{Odds for women} = 0.73 / (1-0.73) = 2.67$$

$$\text{Odds for men} = 0.19 / (1-0.19) = 0.24$$

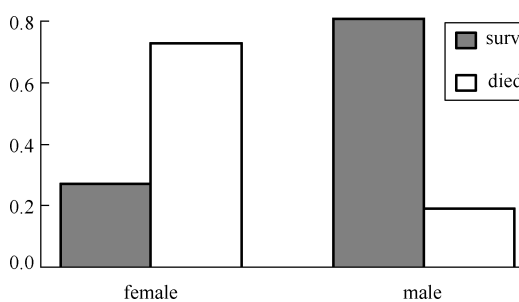


Figure 4.11 19% of men survived vs 73% of women survived

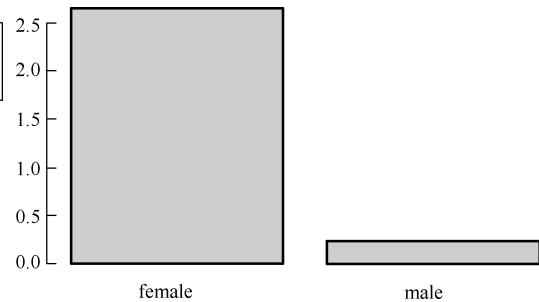


Figure 4.12 Odds for women vs Odds for men

The odds of survival for women is 2.67 and the odds of survival for men is 0.24, see Figure 4.12. The odds ratio (OR) is given by:

$$OR = 2.67/0.24 = 11$$

In English: “Women were eleven times as likely as men to survive the Titanic.”

The odds ratios are constant here, but not the ratios of the probabilities directly. A doubling of

the odds when the odds is already high doesn't have nearly the effect on the probability of doubling the odds when the odds is low. The relationship is reversed if we halve the odds.

Problems

4.1 Find the union $C_1 \cup C_2$ and the intersection $C_1 \cap C_2$ of the two sets C_1 and C_2 , where:

a. $C_1 = \{0, 1, 2\}$, $C_2 = \{2, 3, 4\}$.

b. $C_1 = \{x: 0 < x < 2\}$, $C_2 = \{x: 1 \leq x < 3\}$.

c. $C_1 = \{(x, y): 0 < x < 2, 1 < y < 2\}$, $C_2 = \{(x, y): 1 < x < 3, 1 < y < 3\}$.

4.2 You and I play a coin-tossing game: if the coin falls heads I score one, if tails you score one. In the beginning, the score is zero.

a. What is the probability that after $2n$ throws our scores are equal?

b. What is the probability that after $2n+1$ throws my score is three more than yours?

4.3 There are n people gathered in a room.

a. What is the probability that two (at least) have the same birthday? Calculate the probability for $n = 22$ and 23 .

b. What is the probability that at least one has the same birthday as you? What value of n makes it close to $1/2$?

4.4 In a poker hand consisting of 5 cards, find the probability of holding 2 ace and 3 jacks.

4.5 A and B are events defined on sample space, with $P(A) = 0.7$ and $P(B|A) = 0.4$. Find $P(A \text{ and } B)$.

4.6 A and B are events defined on sample space, with $P(A) = 0.6$ and $P(A \text{ and } B) = 0.3$. Find $P(B|A)$.

4.7 Given $P(A \text{ or } B) = 1.0$, $P(\overline{A \text{ and } B}) = 0.3$, and $P(\overline{B}) = 0.4$. Find:

a. $P(B)$

b. $P(A)$

c. $P(A|B)$

4.8 Determine whether each of the following pairs of event is mutually exclusive.

a. Five coins are tossed: "one head is observed", "at least one head is observed".

b. A salesperson calls on a client and makes a sale: "the sale exceeds \$100", "the sale exceeds \$1000."

c. One student is selected at random from a student body: the person selected is "male", the person selected is "older than 21 years of age".

d. Two dices are rolled: the total showing is "less than 7", the total showing is "more than 9".

4.9 A and B are independent events, and $P(A) = 0.7$, and $P(B) = 0.4$. Find $P(A \text{ and } B)$.

4.10 $P(R) = 0.5$, $P(S) = 0.4$, and R and S are independent.

a. Find $P(R \text{ and } S)$.

b. Find $P(R \text{ or } S)$.

c. $P(\overline{S})$.

d. Find $P(R | S)$.

e. Find $P(\overline{S} | R)$.

Probability theory is nothing but common sense reduced to calculation.

—— Pierre-Simon Laplace (1749—1827), French mathematician

Note: Define independence in terms of conditional probability, and test for independence in that manner.



Unit 5

Discrete Probability Models



5.1 Introduction, Mass Functions and Distribution Functions



5.2 Expectation and Variance for Discrete Random Quantities



5.3 Properties of Expectation and Variance



5.4 The Binomial Distribution



5.5 The Geometric Distribution



5.6 The Poisson Distribution



Reading English Materials



Problems

5.1 Introduction, Mass Functions and Distribution Functions

5.1.1 Introduction

We now have a good understanding of basic probabilistic reasoning. We have seen how to relate events to sets, and how to calculate probabilities for events by working with the sets that represent them. So far, however, we haven't developed any special techniques for thinking about *random quantities*. *Discrete probability models* provide a framework for thinking about *discrete random quantities*, and *continuous probability models* (to be considered in the next unit) form a framework for thinking about *continuous random quantities*.

Example 5.1

Consider the sample space for tossing a fair coin twice:

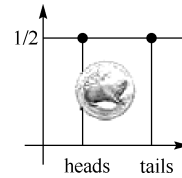
$$S = \{HH, HT, TH, TT\}.$$

These outcomes are equally likely. There are several random quantities we could associate with this experiment. For example, we could count the number of heads, or the number of tails.

Formally, a *random quantity* is a real valued function which acts on *elements* of the sample space (outcomes). That is, to each outcome, the random variable assigns a real number. Random quantities (sometimes known as *random variables*) are always denoted by upper case letters.

In our example, if we let X be the number of heads, we have

$$\begin{aligned} X(HH) &= 2, \\ X(HT) &= 1, \\ X(TH) &= 1, \\ X(TT) &= 0. \end{aligned}$$



The observed value of a random quantity is the number corresponding to the actual outcome. That is, if the outcome of an experiment is $s \in S$, then $X(s) \in \mathbb{R}$ is the observed value. This observed value is always denoted with a lower case letter — here x . Thus $X = x$ means that the observed value of the random quantity, X is the number x . The set of possible observed values for X is

$$S_X = \{X(s) \mid s \in S\}.$$

For the above example we have

$$S_X = \{0, 1, 2\}.$$

Clearly here the values are not all equally likely.

Example 5.2

Roll one die and call the random number which is uppermost Y . The sample space for the *random quantity* Y is

$$S_Y = \{1, 2, 3, 4, 5, 6\}$$

and these outcomes are all equally likely, see Figure 5.1.

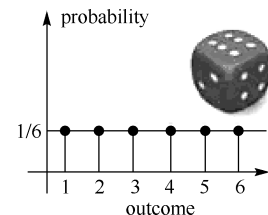


Figure 5.1 Probability and outcome

Now roll two dice and call their sum Z . The sample space for Z is

$$S_Z = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$$

and these outcomes are *not* equally likely. However, we know the probabilities of the events corresponding to each of these outcomes, and we could display them in a table as follows.

Outcome	2	3	4	5	6	7	8	9	10	11	12
Probability	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

This is essentially a tabulation of the *probability mass function* for the random quantity Z .

5.1.2 Probability Mass Functions (PMFs)

For any discrete random variable X , we define the *probability mass function* (PMF) to be the function which gives the probability of each $x \in S_X$. Clearly we have

$$P(X = x) = \sum_{\{s \in S | X(s) = x\}} P(\{s\}).$$

That is, the probability of getting a particular number is the sum of the probabilities of all those outcomes which have that number associated with them. Also $P(X = x) \geq 0$ for each $x \in S_X$, and $P(X = x) = 0$ otherwise. The set of all pairs $\{(x, P(X = x)) | x \in S_X\}$ is known as the *probability distribution* of X .

Definition 1

■ **Probability Distribution:** The set of all pairs $\{(x, P(X = x)) | x \in S_X\}$ is known as the *probability distribution* of X .

Example 5.3

For the example above concerning the sum of two dice, the probability distribution is $\{(2, 1/36), (3, 2/36), (4, 3/36), (5, 4/36), (6, 5/36), (7, 6/36), (8, 5/36), (9, 4/36), (10, 3/36), (11, 2/36), (12, 1/36)\}$ and the probability mass function can be tabulated as

x	2	3	4	5	6	7	8	9	10	11	12
$P(X = x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

and plotted graphically as follows, see Figure 5.2.

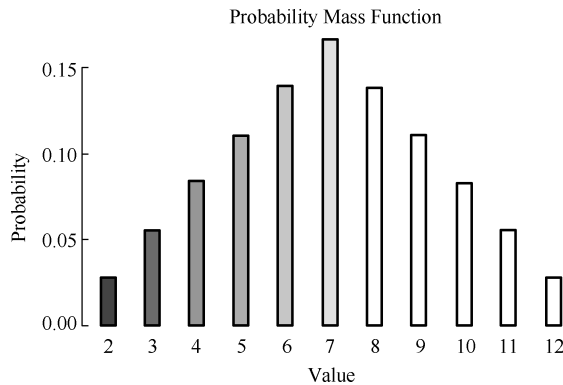


Figure 5.2 Probability mass function

5.1.3 Cumulative Distribution Functions (CDFs)

For any discrete random quantity, X , we clearly have

$$\sum_{x \in S_X} P(X = x) = 1$$

as every outcome has some number associated with it. It can often be useful to know the probability that your random number is no greater than some particular value. With that in mind, we define the *cumulative distribution function*,

$$F_X(x) = P(X \leq x) = \sum_{\{y \in S_X | y \leq x\}} P(X = y).$$

Definition 2

■ Cumulative Distribution Function:

$$F_X(x) = P(X \leq x) = \sum_{\{y \in S_X | y \leq x\}} P(X = y).$$

Example 5.4

For the sum of two dice, the CDF can be tabulated for the outcomes as

x	2	3	4	5	6	7	8	9	10	11	12
$F_X(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

but it is important to note that the CDF is defined *for all real numbers* — not just the possible values. In our example we have

$$\begin{aligned} F_X(-3) &= P(X \leq -3) = 0, \\ F_X(4.5) &= P(X \leq 4.5) = P(X \leq 4) = 6/36, \\ F_X(25) &= P(X \leq 25) = 1. \end{aligned}$$

We may plot the CDF for our example as follows, see Figure 5.3.

It is clear that for any random variable X , for all $x \in \mathbb{R}$, $F_X(x) \in [0, 1]$ and that $F_X(x) \rightarrow 0$ as $x \rightarrow -\infty$ and $F_X(x) \rightarrow 1$ as $x \rightarrow +\infty$.

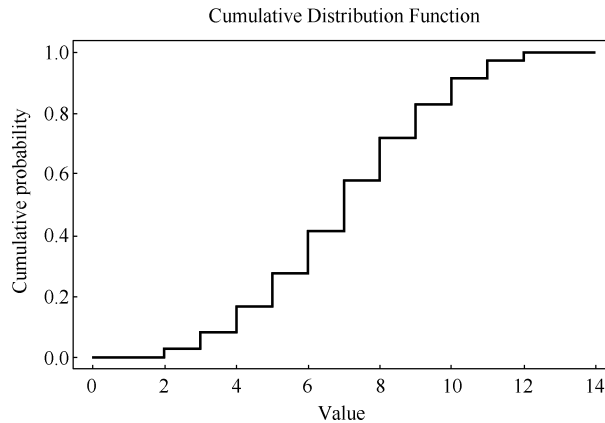


Figure 5.3 Cumulative distribution function

New Words and Expressions

framework ['freɪmwɜ:k] *n.* 构架; 框架; (体系的) 结构

uppermost ['ʌpəməʊst] *adj.* 最高的; 至上的; 最重要的

dice [daɪs] *n.* 骰子; 掷骰游戏

Technical Terms

probability mass function 概率质量函数, 记为 PMF 或 PMFs

cumulative distribution function 累计分布函数, 记为 CDF 或 CDFs

Notes

1. upper case letter 也常写成 uppercase letter, 表示大写字母。类似地 lowercase letter 表示小写字母

5.2 Expectation and Variance for Discrete Random Quantities

5.2.1 Expectation

Just as it is useful to summarize data (see Unit 1), it is just as useful to be able to summarize the distribution of random quantities. The *location measure* used to summarize random quantities is known as the *expectation* of the random quantity. It is the “centre of mass” of the probability distribution. The expectation of a discrete random quantity X , written $E(X)$ is defined by

$$E(X) = \sum_{x \in \mathcal{S}_X} x P(X = x).$$

The expectation is often denoted by $E(X)$ or even just μ . Note that the expectation is a known function of the probability distribution. It is *not* a random quantity, and in particular, it is *not* the sample mean of a set of data (random or otherwise). In fact, there is a *relationship* between the sample mean of a set of data and the expectation of the underlying probability distribution generating the data.

Example 5.5

For the sum of two dice, X , we have

$$E(X) = 2 \times \frac{1}{36} + 3 \times \frac{2}{36} + 4 \times \frac{3}{36} + \cdots + 12 \times \frac{1}{36} = 7.$$

By looking at the symmetry of the mass function, it is clear that in some sense 7 is the “central” value of the probability distribution.

5.2.2 Variance

We now have a method for summarizing the location of a given probability distribution, but we also need a summary for the *spread*.

Definition 3

■ For a Discrete Random Quantity X , the *variance* of X is defined by

$$\text{Var}(X) = \sum_{x \in S_X} \{(x - E(X))^2 P(X=x)\}.$$

For a discrete random quantity X , the *variance* of X is defined by

$$\text{Var}(X) = \sum_{x \in S_X} \{(x - E(X))^2 P(X=x)\}.$$

The variance is often denoted σ_X^2 , or even just σ^2 . Again, this is a known function of the probability distribution. It is not random, and it is not the *sample* variance of a set of data. Again, the two are related in a way to be made precise later. The variance can be rewritten as

$$\text{Var}(X) = \sum_{x_i \in S_X} x_i^2 P(X=x_i) - [E(X)]^2,$$

and this expression is usually a bit easier to work with. We also define the *standard deviation* of a random quantity by

$$\text{SD}(X) = \sqrt{\text{Var}(X)},$$

and this is usually denoted by σ_X or just σ .

Example 5.6

For the sum of two dice, X , we have

$$\sum_{x_i \in S_X} x_i^2 P(X=x_i) = 2^2 \times \frac{1}{36} + 3^2 \times \frac{2}{36} + 4^2 \times \frac{3}{36} + \cdots + 12^2 \times \frac{1}{36} = \frac{329}{6}$$

and so

$$\text{Var}(X) = \frac{329}{6} - 7^2 = \frac{35}{6},$$

and

$$\text{SD}(X) = \sqrt{\frac{35}{6}}.$$

New Words and Expressions

spread [spred] *n.* 范围; 连续的一段时间; [统计] 散布

Technical Terms

location measure 位置测量

standard deviation 标准差[GB]

5.3 Properties of Expectation and Variance

One of the reasons that expectation is widely used as a measure of location for probability distributions is the fact that it has many desirable mathematical properties which make it elegant and convenient to work with. Indeed, many of the nice properties of expectation lead to corresponding nice properties for variance, which is one of the reasons why variance is widely used as a measure of spread.

5.3.1 Expectation of a Function of a Random Quantity

Suppose that X is a discrete random quantity, and that Y is another random quantity that is a known function of X . That is, $Y = g(X)$ for some function $g(\cdot)$. What is the expectation of Y ?

Example 5.7

Throw a die, and let X be the number showing. We have

$$S_X = \{1, 2, 3, 4, 5, 6\}$$

and each value is equally likely. Now suppose that we are actually interested in the square of the number showing. Define a new random quantity $Y = X^2$. Then

$$S_Y = \{1, 4, 9, 16, 25, 36\}$$

and clearly each of these values is equally likely. We therefore have

$$E(Y) = 1 \times \frac{1}{6} + 4 \times \frac{1}{6} + \cdots + 36 \times \frac{1}{6} = \frac{91}{6}.$$

The above example illustrates the more general result, that for $Y = g(X)$, we have

$$E(Y) = \sum_{x \in S_X} g(x) P(X = x).$$

Note that *in general* $E(g(X)) \neq g(E(X))$. For the above example, $E(X^2) = 91/6 \approx 15.2$, and $E(X)^2 = 3.5^2 = 12.25$.

We can use this more general notion of expectation in order to redefine variance purely in terms of expectation as follows:

$$\text{Var}(X) = E([X - E(X)]^2) = E(X^2) - E(X)^2.$$

Having said that $E(g(X)) \neq g(E(X))$ in general, it does in fact hold in the (very) special, but important case where $g(\cdot)$ is a *linear* function.

5.3.2 Expectation of a Linear Transformation

◇ Expectation of a Linear Transformation ◇

If we have a random quantity X , and a linear transformation, $Y = aX + b$, where a and b are known real constants, then we have that

$$E(aX + b) = aE(X) + b.$$

We can show this as follows:

$$\begin{aligned}
E(aX + b) &= \sum_{x \in S_X} (ax + b) P(X = x) \\
&= \sum_{x \in S_X} ax P(X = x) + \sum_{x \in S_X} b P(X = x) \\
&= a \sum_{x \in S_X} x P(X = x) + b \sum_{x \in S_X} P(X = x) \\
&= a E(X) + b.
\end{aligned}$$

5.3.3 Expectation of the Sum of Two Random Quantities

For two random quantities X and Y , the expectation of their sum is given by

$$E(X + Y) = E(X) + E(Y).$$

Note that this result is true irrespective of whether or not X and Y are independent. Let us see why. First,

$$S_{X+Y} = \{x + y | (x \in S_X) \cap (y \in S_Y)\},$$

and so

$$\begin{aligned}
E(X + Y) &= \sum_{(x+y) \in S_{X+Y}} (x + y) P((X = x) \cap (Y = y)) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} (x + y) P((X = x) \cap (Y = y)) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} x P((X = x) \cap (Y = y)) \\
&\quad + \sum_{x \in S_X} \sum_{y \in S_Y} y P((X = x) \cap (Y = y)) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} x P(X = x) P(Y = y | X = x) \\
&\quad + \sum_{y \in S_Y} \sum_{x \in S_X} y P(Y = y) P(X = x | Y = y) \\
&= \sum_{x \in S_X} x P(X = x) \sum_{y \in S_Y} P(Y = y | X = x) \\
&\quad + \sum_{y \in S_Y} y P(Y = y) \sum_{x \in S_X} P(X = x | Y = y) \\
&= \sum_{x \in S_X} x P(X = x) + \sum_{y \in S_Y} y P(Y = y) \\
&= E(X) + E(Y).
\end{aligned}$$

5.3.4 Expectation of an Independent Product

If X and Y are *independent* random quantities, then

$$E(XY) = E(X) E(Y).$$

To see why, note that

$$S_{XY} = \{xy | (x \in S_X) \cap (y \in S_Y)\},$$

and so

$$\begin{aligned}
E(XY) &= \sum_{xy \in S_{XY}} xy P((X=x) \cap (Y=y)) \\
&= \sum_{x \in S_X} \sum_{y \in S_Y} xy P(X=x) P(Y=y) \\
&= \sum_{x \in S_X} x P(X=x) \sum_{y \in S_Y} y P(Y=y) \\
&= E(X) E(Y)
\end{aligned}$$

Note that here it is *vital* that X and Y are independent, or the result does not hold.

5.3.5 Variance of an Independent Sum

If X and Y are *independent* random quantities, then

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y).$$

To see this, write

$$\begin{aligned}
\text{Var}(X+Y) &= E([X+Y]^2) - [E(X+Y)]^2 \\
&= E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2 \\
&= E(X^2) + 2E(XY) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\
&= E(X^2) + 2E(X)E(Y) + E(Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\
&= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 \\
&= \text{Var}(X) + \text{Var}(Y)
\end{aligned}$$

Again, it is vital that X and Y are independent, or the result does not hold. Notice that this implies a slightly less attractive result for the standard deviation of the sum of two independent random quantities,

$$\text{SD}(X+Y) = \sqrt{\text{SD}(X)^2 + \text{SD}(Y)^2},$$

which is why it is often more convenient to work with variances.

New Words and Expressions

elegant ['elɪɡənt] *adj.* (人或其举止) 优美的; 漂亮的; 简洁的

vital ['vaɪtl] *adj.* 至关重要的; 生死攸关的; 生气勃勃的

5.4 The Binomial Distribution

5.4.1 Introduction

Now that we have a good understanding of discrete random quantities and their properties, we can go on to look at a few of the standard families of discrete random variables. One of the most commonly encountered discrete distributions is the binomial distribution. This is the distribution of the number of “successes” in a series of independent “success”/“fail” trials. Before we look at this, we need to make sure we understand the case of a single trial.

5.4.2 Bernoulli Random Quantities

Suppose that we have an event E in which we are interested, and we write its sample space as

$$S = \{E, E^c\}.$$

We can associate a random quantity with this sample space, traditionally denoted I , as $I(E) = 1$, $I(E^c) = 0$. So, if $P(E) = p$, we have

$$S_I = \{0, 1\},$$

and $P(I = 1) = p$, $P(I = 0) = 1 - p$. This random quantity, I is known as an *indicator variable*, and is often useful for constructing more complex random quantities. We write

$$I \sim \text{Bern}(p).$$

We can calculate its expectation and variance as follows.

$$\begin{aligned} E(I) &= 0 \times (1 - p) + 1 \times p = p \\ E(I^2) &= 0^2 \times (1 - p) + 1^2 \times p = p \\ \text{Var}(I) &= E(I^2) - E(I)^2 \\ &= p - p^2 = p(1 - p) \end{aligned}$$

With these results, we can now go on to understand the binomial distribution.

5.4.3 The Binomial Distribution

The binomial distribution is the distribution of the number of “successes” in a series of n independent “trials”, each of which results in a “success” (with probability p) or a “failure” (with probability $1 - p$). If the number of successes is X , we would write

$$X \sim B(n, p)$$

to indicate that X is a binomial random quantity based on n independent trials, each occurring with probability p .

Example 5.8

(1) Toss a fair coin 100 times and let X be the number of heads. Then $X \sim B(100, 0.5)$.

(2) A certain kind of lizard lays 8 eggs, each of which will hatch independently with probability 0.7. Let Y denote the number of eggs which hatch. Then $Y \sim B(8, 0.7)$.

Let us now derive the probability mass function for $X \sim B(n, p)$. Clearly X can take on any value from 0 up to n , and no other. Therefore, we simply have to calculate $P(X = k)$ for $k = 0, 1, 2, \dots, n$. The probability of k successes followed by $n - k$ failures is clearly $p^k (1 - p)^{n-k}$. Indeed, this is the probability of *any* particular sequence involving k successes. There are $\binom{n}{k}$ such sequences, so by the multiplication principle, we have

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Now, using the binomial theorem, we have

$$\sum_{k=0}^n P(X=k) = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = (p + [1-p])^n = 1^n = 1,$$

and so this does define a valid probability distribution.

Example 5.9

For the lizard eggs, $Y \sim B(8, 0.7)$ we have

$$P(Y=k) = \binom{8}{k} 0.7^k 0.3^{8-k}, \quad k = 0, 1, 2, \dots, 8.$$

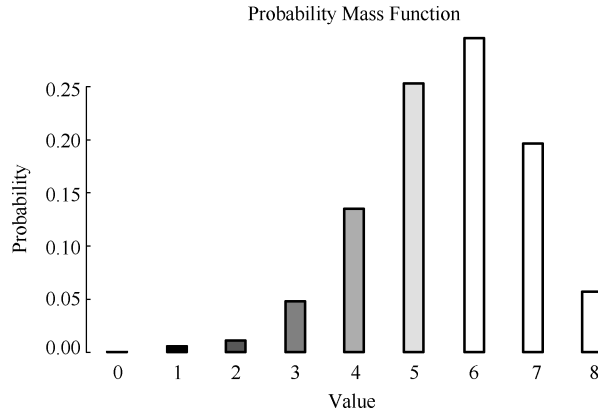


Figure 5.4 Probability mass function

We can therefore tabulate and plot the probability mass function and cumulative distribution function as follows, see Figure 5.4 and see Figure 5.5.

k	0	1	2	3	4	5	6	7	8
$P(Y=k)$	0.00	0.00	0.01	0.05	0.14	0.25	0.30	0.20	0.06
$F_Y(k)$	0.00	0.00	0.01	0.06	0.19	0.45	0.74	0.94	1.00

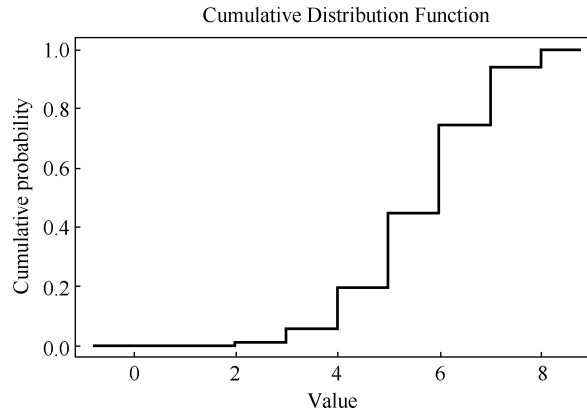


Figure 5.5 Cumulative distribution function

Similarly, the PMF and CDF for $X \sim B(100, 0.5)$ (number of heads from 100 coin tosses) can be plotted as follows, see Figure 5.6 and see Figure 5.7.

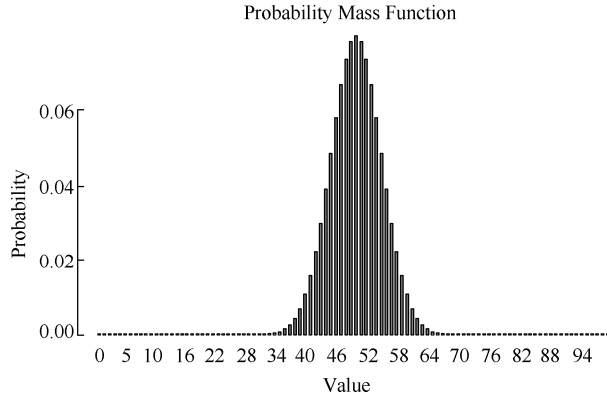


Figure 5.6 Probability mass functions

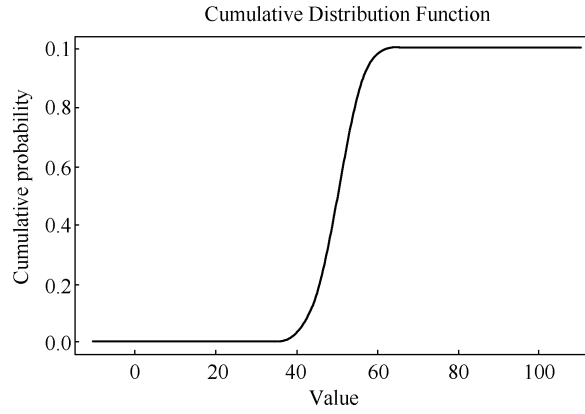


Figure 5.7 Cumulative distribution function

5.4.4 Expectation and Variance of a Binomial Random Quantity

It is possible (but a little messy) to derive the expectation and variance of the binomial distribution directly from the PMF. However, we can deduce them rather more elegantly if we recognize the relationship between the binomial and Bernoulli distributions. If $X \sim B(n, p)$ then

$$X = \sum_{j=1}^n I_j$$

where $I_j \sim \text{Bern}(p)$, $j = 1, 2, \dots, n$, and the I_j are mutually independent. So we then have

$$\begin{aligned} E(X) &= E\left(\sum_{j=1}^n I_j\right) \\ &= \sum_{j=1}^n E(I_j) && \text{(expectation of a sum)} \\ &= \sum_{j=1}^n p \\ &= np \end{aligned}$$

and similarly,

$$\begin{aligned}
\text{Var}(X) &= \text{Var}\left(\sum_{j=1}^n I_j\right) \\
&= \sum_{j=1}^n \text{Var}(I_j) && \text{(variance of independent sum)} \\
&= \sum_{j=1}^n p(1-p) \\
&= np(1-p).
\end{aligned}$$

Example 5.10

For the coin tosses, $X \sim B(100, 0.5)$,

$$\begin{aligned}
E(X) &= np = 100 \times 0.5 = 50, \\
\text{Var}(X) &= np(1-p) = 100 \times 0.5^2 = 25,
\end{aligned}$$

and so

$$\text{SD}(X) = 5.$$

Similarly, for the lizard eggs, $Y \sim B(8, 0.7)$,

$$\begin{aligned}
E(Y) &= np = 8 \times 0.7 = 5.6, \\
\text{Var}(Y) &= np(1-p) = 8 \times 0.7 \times 0.3 = 1.68
\end{aligned}$$

and so

$$\text{SD}(Y) = 1.30.$$

New Words and Expressions

lizard ['lɪzəd] *n.* 蜥蜴

deduce [dɪ'dju:s] *vt.* 推论, 推断; 演绎。~+[from]演绎, 推断, 推论

Technical Terms

indicator variable 指示变量

5.5 The Geometric Distribution

5.5.1 PMF

The geometric distribution is the distribution of the number of independent Bernoulli trials until the first success is encountered. If X is the number of trials until a success is encountered, and each independent trial has probability p of being a success, we write

$$X \sim \text{Geom}(p).$$

Clearly X can take on any positive integer, so to deduce the PMF, we need to calculate $P(X = k)$ for $k = 1, 2, 3, \dots$. In order to have $X = k$, we must have an ordered sequence of $k - 1$ failures followed by one success. By the multiplication rule therefore,

$$P(X = k) = (1 - p)^{k-1} p, \quad k = 1, 2, 3, \dots$$

5.5.2 CDF

For the geometric distribution, it is possible to calculate an analytic form for the CDF as follows. If $X \sim \text{Geom}(p)$, then

$$\begin{aligned}
 F_X(k) &= P(X \leq k) \\
 &= \sum_{j=1}^k (1-p)^{j-1} p \\
 &= p \sum_{j=1}^k (1-p)^{j-1} \\
 &= p \cdot \frac{1 - (1-p)^k}{1 - (1-p)} \quad (\text{geometric series}) \\
 &= 1 - (1-p)^k.
 \end{aligned}$$

Consequently there is no need to tabulate the CDF of the geometric distribution. Also note that the CDF tends to one as k increases. This confirms that the PMF we defined does determine a valid probability distribution.

Example 5.11

Suppose that we are interested in playing a game where the probability of winning is 0.2 on any particular turn. If X is the number of turns until the first win, then $X \sim \text{Geom}(0.2)$. The PMF and CDF for X are plotted below, see Figure 5.8 and see Figure 5.9.

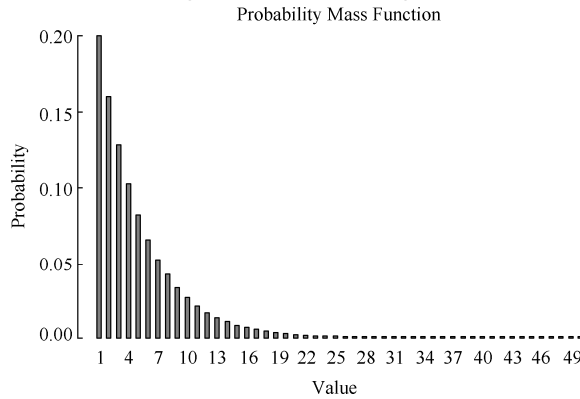


Figure 5.8 Probability mass functions

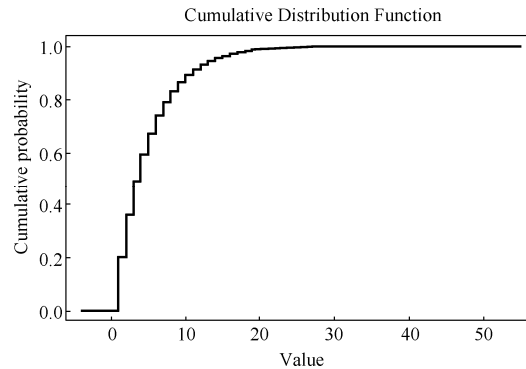


Figure 5.9 Cumulative distribution function

5.5.3 Useful Series in Probability

Notice that we used the sum of a geometric series in the derivation of the CDF. There are many other series that crop up in the study of probability. A few of the more commonly encountered series are listed below.

$$\begin{aligned}\sum_{i=1}^n a^{i-1} &= \frac{1-a^n}{1-a} & (a > 0) \\ \sum_{i=1}^{\infty} a^{i-1} &= \frac{1}{1-a} & (0 < a < 1) \\ \sum_{i=1}^{\infty} i a^{i-1} &= \frac{1}{(1-a)^2} & (0 < a < 1) \\ \sum_{i=1}^{\infty} i^2 a^{i-1} &= \frac{1+a}{(1-a)^3} & (0 < a < 1) \\ \sum_{i=1}^n i &= \frac{n(n+1)}{2} \\ \sum_{i=1}^n i^2 &= \frac{1}{6}n(n+1)(2n+1).\end{aligned}$$

We will use two of these in the derivation of the expectation and variance of the geometric distribution.

5.5.4 Expectation and Variance of Geometric Random Quantities

Suppose that $X \sim \text{Geom}(p)$. Then

$$\begin{aligned}E(X) &= \sum_{i=1}^{\infty} i P(X=i) \\ &= \sum_{i=1}^{\infty} i(1-p)^{i-1} p \\ &= p \sum_{i=1}^{\infty} i(1-p)^{i-1} \\ &= p \cdot \frac{1}{(1-[1-p])^2} \\ &= \frac{p}{p^2} \\ &= \frac{1}{p}.\end{aligned}$$

Similarly,

$$\begin{aligned}E(X^2) &= \sum_{i=1}^{\infty} i^2 P(X=i) \\ &= \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} p \\ &= p \sum_{i=1}^{\infty} i^2 (1-p)^{i-1} \\ &= p \cdot \frac{1+[1-p]}{(1-[1-p])^3} \\ &= p \cdot \frac{2-p}{p^3} \\ &= \frac{2-p}{p^2},\end{aligned}$$

and so

$$\begin{aligned}\text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}.\end{aligned}$$

Example 5.12

For $X \sim \text{Geom}(0.2)$ we have

$$\begin{aligned}E(X) &= \frac{1}{p} = \frac{1}{0.2} = 5 \\ \text{Var}(X) &= \frac{1-p}{p^2} = \frac{0.8}{0.2^2} = 20.\end{aligned}$$

New Words and Expressions

crop [krɒp] *vt.* 种植; 收割; 修剪 *vi.* 收成; 收获
crop up 显露出来, 发生, 出现

Technical Terms

geometric distribution 几何分布
geometric series 几何序列

5.6 The Poisson Distribution

The Poisson distribution is a very important discrete probability distribution, which arises in many different contexts in probability and statistics. Typically, Poisson random quantities are used in place of binomial random quantities in situations where n is large, p is small, and the expectation np is stable.

Example 5.13

Consider the number of calls made in a 1 minute interval to an Internet service provider (ISP). The ISP has thousands of subscribers, but each one will call with a very small probability. The ISP knows that on average 5 calls will be made in the interval. The actual number of calls will be a Poisson random variable, with mean 5.

A Poisson random variable, X with parameter λ is written as

$$X \sim P(\lambda)$$

5.6.1 Poisson as the Limit of a Binomial

Let $X \sim B(n, p)$. Put $\lambda = E(X) = np$ and let n increase and p decrease so that λ remains constant.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \dots, n.$$

Replacing p by λ/n gives

$$\begin{aligned}
P(X = k) &= \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \frac{n!}{(n-k)!n^k} \frac{(1 - \lambda/n)^n}{(1 - \lambda/n)^k} \\
&= \frac{\lambda^k}{k!} \frac{n}{n} \frac{(n-1)}{n} \frac{(n-2)}{n} \dots \frac{(n-k+1)}{n} \frac{(1 - \lambda/n)^n}{(1 - \lambda/n)^k} \\
&\rightarrow \frac{\lambda^k}{k!} \times 1 \times 1 \times 1 \times \dots \times 1 \times \frac{e^{-\lambda}}{1}, \quad \text{as } n \rightarrow \infty \\
&= \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}$$

To see the limit, note that $(1 - \lambda/n)^n \rightarrow e^{-\lambda}$ as n increases (compound interest formula).

5.6.2 PMF

If $X \sim P(\lambda)$, then the PMF of X is

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, 3, \dots$$

Example 5.14

The PMF and CDF of $X \sim P(5)$ are given below, see Figure 5.10 and see Figure 5.11.

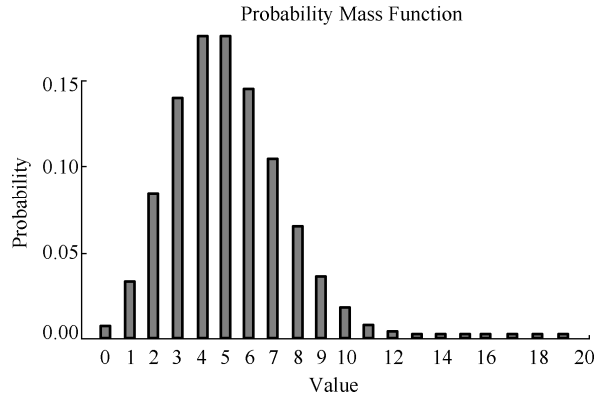


Figure 5.10 Probability mass function

Note that the CDF does seem to tend to 1 as n increases. However, we do need to verify that the PMF we have adopted for $X \sim P(\lambda)$ does indeed define a valid probability distribution, by ensuring that the probabilities do sum to one, see Figure 5.12.

$$\begin{aligned}
P(S_X) &= \sum_{k=0}^{\infty} P(X = k) \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.
\end{aligned}$$

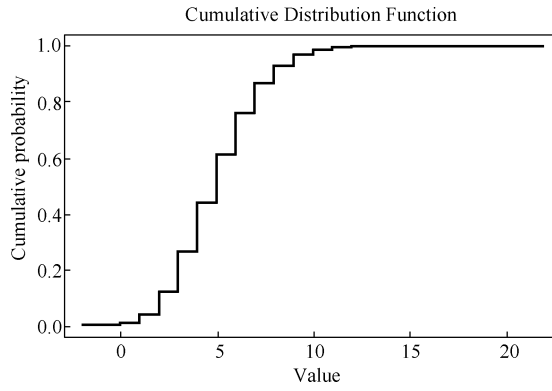


Figure 5.11 Cumulative distribution function

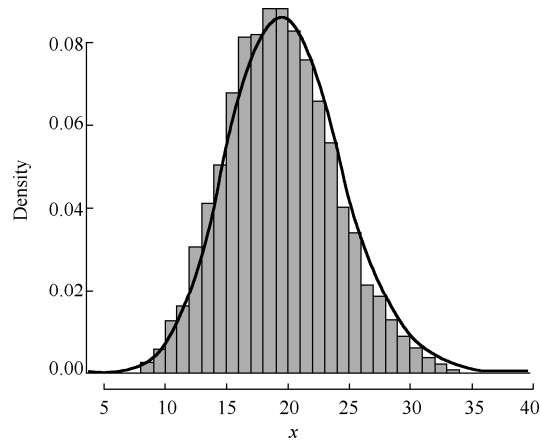


Figure 5.12 Poisson distribution, here lambda = 20.

5.6.3 Expectation and Variance of Poisson

If $X \sim P(\lambda)$, we have

$$\begin{aligned}
 E(X) &= \sum_{k=0}^{\infty} k P(X = k) \\
 &= \sum_{k=1}^{\infty} k P(X = k) \\
 &= \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\
 &= \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} e^{-\lambda} \\
 &= \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
 &= \lambda \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} \quad (\text{putting } j = k - 1) \\
 &= \lambda \sum_{j=0}^{\infty} P(X = j) \\
 &= \lambda.
 \end{aligned}$$

Similarly,

$$\begin{aligned}
E(X^2) &= \sum_{k=0}^{\infty} k^2 P(X=k) \\
&= \sum_{k=1}^{\infty} k^2 P(X=k) \\
&= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} \\
&= \lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} \\
&= \lambda \sum_{j=0}^{\infty} (j+1) \frac{\lambda^j}{j!} e^{-\lambda} && \text{(putting } j = k-1) \\
&= \lambda \left[\sum_{j=0}^{\infty} j \frac{\lambda^j}{j!} e^{-\lambda} + \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} e^{-\lambda} \right] \\
&= \lambda \left[\sum_{j=0}^{\infty} j P(X=j) + \sum_{j=0}^{\infty} P(X=j) \right] \\
&= \lambda [E(X) + 1] \\
&= \lambda(\lambda + 1) \\
&= \lambda^2 + \lambda.
\end{aligned}$$

So,

$$\begin{aligned}
\text{Var}(X) &= E(X^2) - E(X)^2 \\
&= [\lambda^2 + \lambda] - \lambda^2 \\
&= \lambda.
\end{aligned}$$

That is, the mean and variance are both λ .

5.6.4 Sum of Poisson Random Quantities

One of the particularly convenient properties of the Poisson distribution is that the sum of two independent Poisson random quantities is also a Poisson random quantity. If $X \sim P(\lambda)$ and $Y \sim P(\mu)$ and X and Y are independent, then $Z = X + Y \sim P(\lambda + \mu)$. Clearly this result extends to the sum of many independent Poisson random variables. The proof is straightforward, but is a little messy, and hence omitted from this course.

Example 5.15

Returning to the example of calls received by an ISP. The number of calls in 1 minute is $X \sim P(5)$. Suppose that the number of calls in the following minute is $Y \sim P(5)$, and that Y is independent of X . Then, by the above result, $Z = X + Y$, the number of calls in the two minute period is Poisson with parameter 10. Extending this in the natural way, we see that the number of calls in t minutes is Poisson with parameter $5t$. This motivates the following definition.

5.6.5 The Poisson Process

A sequence of timed observations is said to follow a *Poisson process* with *rate* λ if the

number of observations, X , in any interval of length t is such that

$$X \sim P(\lambda t).$$

Example 5.16

For the ISP example, the sequence of incoming calls follow a Poisson process with rate 5 (per minute).

New Words and Expressions

subscribers [səbsk'raɪbəz] *n.* (报刊的) 订阅人(subscriber 的名词复数); [英] (慈善机关等的) 定期捐款者; 消费者; 用户

verify ['verɪfaɪ] *vt.* 核实; 证明; 判定

messy ['mesi] *adj.* 凌乱的, 散乱的; 肮脏的, 污秽的; 复杂的, 难以应付的

Technical Terms

Internet service provider (ISP) 互联网服务供应者 (或供应商)

Passage 1. The Founder of Modern Statistics—Karl Pearson

Karl Pearson (1857. 3, 27—1936. 4, 27) originally named Carl, is widely regarded as the founder of the modern discipline of statistics, and is also famous as a philosopher of science, as a writer on social Darwinism and as a leading mover to install eugenics as the key social science.

Karl Pearson as the British mathematician and one of the early contributors to statistical theory and methods, was the first to perceive the use of random numbers for solving problems in probability and statistics that are too complex for exact solution.

In fact, before the beginning of 20th century, statistics meant observed data and descriptive summary figures, such as means, variances, indices, etc., computed from data. With the introduction of the χ^2 test for goodness of fit (specification) by Karl Pearson (1900) and the t test by Gosset (Student, 1908) for drawing inference on the mean of a normal population, statistics started acquiring new meaning as a method of processing data to determine the amount of uncertainty in various generalizations we may make from observed data (sample) to the source of the data (population).

The major steps that led to the establishment and recognition of statistics as a separate scientific discipline and an inevitable tool in improving natural knowledge were made by R. A. Fisher during the decade 1915—1925. Most of the concepts and methods introduced by Fisher are fundamental and continue to provide the framework for the discussion of statistical theory.

Fisher is the author of about 300 research publications (reproduced in 5 volumes of his collected papers) and six books, of which four are on statistics and two on genetics. The originality of his papers, their thought-provoking contents, and many suggestions for further development should, in spite of the lack of mathematical rigor of some of his contributions, provide a stimulus and challenge to research workers for many years to come.

Fisher's work is monumental, both in richness and variety of ideas, and provided the inspiration for phenomenal developments in statistical methodology for applications in all areas of human endeavor during the last 75 years. Some of Fisher's pioneering works have raised bitter controversies that still continue. These controversies have indeed helped in highlighting the intrinsic difficulties in inductive reasoning and seeking refinements in statistical methodology.

The recognition of statistics as a separate scientific discipline came only after the theoretical foundations of the subject were laid and its applications to scientific research was demonstrated by Fisher.

Passage 2. The Relations of Several Discrete Probability Models

Some discrete random quantities (that is random variables) that arise in some applications,

together with their probability mass functions, parameters, and supports, we summarize as follow, see Table 5.1 and Figure 5.13.

Table 5.1 PMF and parameter of several discrete probability models

Distribution	$P(x)$	Parameters and Support
Bernoulli	$P(x) = p^x (1-p)^{1-x}$	$0 < p < 1, x = 0, 1$
Binomial	$\binom{n}{x} p^x (1-p)^{1-x}$	$n = 1, 2, \dots$ $0 < p < 1, x = 0, 1, \dots, n$
Geometric	$p (1-p)^{x-1}$	$0 < p < 1, x = 1, 2, \dots$
Poisson	$\frac{e^{-\lambda} \lambda^x}{x!}$	$\lambda > 0, x = 0, 1, \dots$

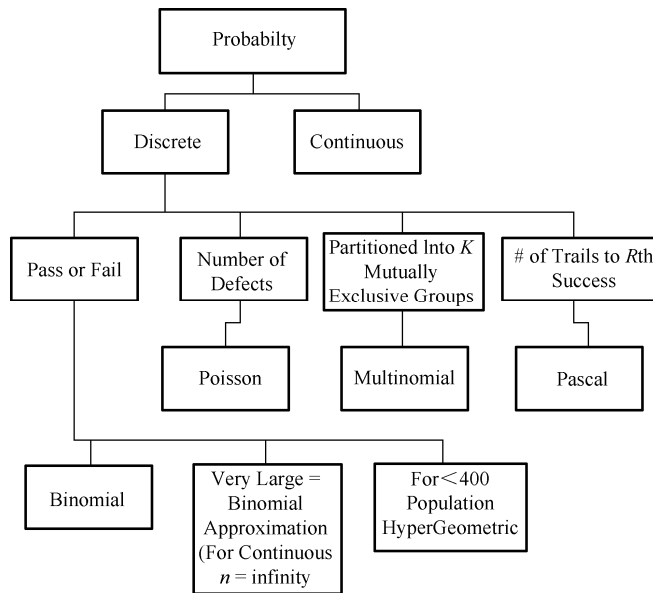


Figure 5.13 The Relations of several discrete probability models

Problems

5.1 Paul and Eric are paying squash(打壁球), and Paul is determined to win at least two games. Unfortunately his chance of winning and one game is only $1/4$, and this chance remains constant however many games he plays against Eric. The players agree to play 5 games and, if Paul has won at least two by then, play ceases. Otherwise Paul persuades(说服) Eric to play a further 5 games with him. What is the probability

- that only 5 games are played, and Paul wins at least two of them.
- that 10 games have to be played, and Paul wins at least two?

5.2 Sampling for Defective Items

A machine produces articles of which an average of 10% are defective. Find an approximate value for the probability that a random sample of 500 of these articles contains more than 25 which are defective. What, approximately, is the probability that the sample contains fewer than 60 defectives?

5.3 The Music Recital(音乐演奏会)

Regular music recital are held a small hall with seating for an audience of 98 people. The booking office staff that, on average, 3% of people who book for recital fail to turn up, and adopt a policy of selling up 100 tickets for any recital. What is the probability that for recital for which 100 tickets have been sold, everyone who turns up has a seat?

5.4 Making a Multiple-choice Test

a. A multiple-choice test paper contains 50 questions; for each question three answers are given, one of which is correct. The two incorrect answers to any question are designed to be plausible, so that an ignorant candidate could be expected to pick an answer quite at random. If the examination is marked simply by giving one mark per correct, what should the pass mark be if the probability that a completely ignorant candidate passes is to be approximately 1%?

b. Suppose now that the examination is marked by awarding two marks per correct answer, but deducting one mark for every incorrect answer. If an ignorant candidate attempts every question, what is the expectation and variance of the candidate's total score?

c. Consider the position of a candidate when the scoring system is as in part (b), but with only one mark for a correct answer, and when the pass mark is 28. The candidate has revised half the syllabus thoroughly, and finds that he is certain of the correct answer to a few more questions. Would the probability of passing be greater if he picked just three more questions hoping to get them all correct, or if guessed at five questions?

5.5 The Telephone Exchange

A telephone exchange receives, on average, 5 calls per minute. Find the probability

- a. that in a 1-minute period no calls are received;
- b. that in a 2-minute period fewer than 4 calls are received;
- c. that in 20-minute period no more than 102 calls are received;
- d. that out of five separate 1-minute periods there are exactly four in which 2 or more calls are received.

5.6 Tossing a Coin until 'Heads' Appears

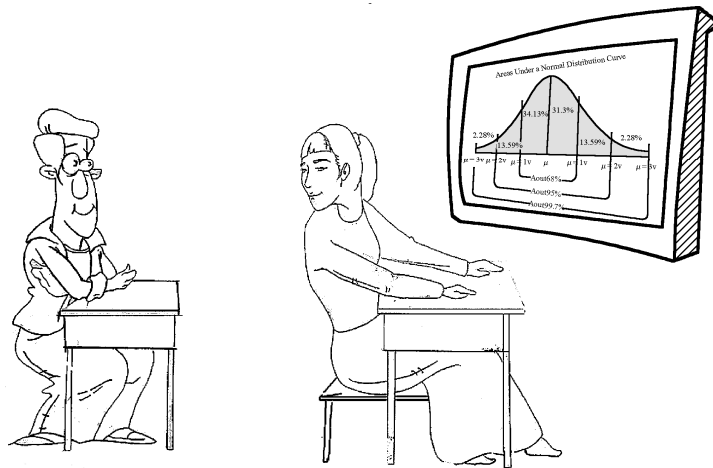
a. In a series of independent tosses of a coin, with probability p of the coin landing 'heads' and probability $1-p$ of it landing 'tails', obtain an expression for $\Pr(X = x)$, $x = 1, 2, 3, \dots$, where X is a random variable denoting the number of tosses until the first head appears.

b. If m is positive integer, obtain an expression for $\Pr(Y = y)$, $y = m, m + 1, m + 2, \dots$, where Y is a random variable denoting the tosses until the m th head appears.

A fundamental goal of statistics is to ensure the reproducibility of scientific findings.

— Victoria Stodden, Professor Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

1. The total area under the standard normal curve is equal 1.
2. The distribution is mound and symmetrical, ...



Unit 6

Discrete Probability Models



6.1 Introduction, PDF and CDF



6.2 Properties of Continuous Random Quantities



6.3 The Uniform Distribution



6.4 The Exponential Distribution



6.5 The Normal Distribution



6.6 The Standard Normal Distribution



6.7 Applications of Normal Distributions



6.8 Specific z -score



6.9 Normal Approximation of Binomial and Poisson



Problems

6.1 Introduction, PDF and CDF

6.1.1 Introduction

We now have a fairly good understanding of discrete probability models, but as yet we haven't developed any techniques for handling continuous random quantities. These are random quantities with a sample space which is neither finite nor countably infinite. The sample space is usually taken to be the real line, or a part thereof. Continuous probability models are appropriate if the result of an experiment is a continuous measurement, rather than a count of a discrete set.

If X is a continuous random quantity with sample space S_X , then for any particular $a \in S_X$, we generally have that

$$P(X = a) = 0.$$

This is because the sample space is so “large” and every possible outcome so “small” that the probability of any *particular* value is vanishingly small. Therefore the probability mass function we defined for discrete random quantities is inappropriate for understanding continuous random quantities. In order to understand continuous random quantities, we need a little calculus.

6.1.2 The Probability Density Function

Definition 1

■ Probability Density Function (PDF):

If X is a *continuous* random quantity, then there exists a function $f_X(x)$, called the *probability density function* (PDF), which satisfies the following:

- (i) $f_X(x) \geq 0, \quad \forall x$;
- (ii) $\int_{-\infty}^{\infty} f_X(x) dx = 1$;
- (iii) $P(a \leq X \leq b) = \int_a^b f_X(x) dx$ for any a and b .

Consequently we have

$$\begin{aligned} P(x \leq X \leq x + \delta x) &= \int_x^{x+\delta x} f_X(y) dy \\ &\simeq f_X(x) \delta x, && \text{(for small } \delta x) \\ \Rightarrow f_X(x) &\simeq \frac{P(x \leq X \leq x + \delta x)}{\delta x} \end{aligned}$$

and so we may interpret the PDF as

$$f_X(x) = \lim_{\delta x \rightarrow 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}.$$

Example 6.1

The manufacturer of a certain kind of light bulb claims that the lifetime of the bulb in hours, X can be modelled as a random quantity with PDF

$$f_X(x) = \begin{cases} 0, & x < 100 \\ \frac{c}{x^2}, & x \geq 100, \end{cases}$$

where c is a constant. What value must c take in order for this to define a valid PDF? What is the probability that the bulb lasts no longer than 150 hours? Given that a bulb lasts longer than 150 hours, what is the probability that it lasts longer than 200 hours?

Notes

(1) Remember that PDFs are *not* probabilities. For example, the density can take values greater than 1 in some regions as long as it still integrates to 1.

(2) It is sometimes helpful to think of a PDF as the limit of a relative frequency histogram for many realizations of the random quantity, where the number of realizations is very large and the bin widths are very small.

(3) Because $P(X = a) = 0$, we have $P(X \leq k) = P(X < k)$ for continuous random quantities.

6.1.3 The Distribution Function

In the last unit 5, we defined the cumulative *distribution function* of a random variable X to be

$$F_X(x) = P(X \leq x), \quad \forall x.$$

This definition works just as well for continuous random quantities, and is one of the many reasons why the distribution function is so useful. For a discrete random quantity we had

$$F_X(x) = P(X \leq x) = \sum_{\{y \in S_X | y \leq x\}} P(X = y),$$

but for a continuous random quantity we have the continuous analogue

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(-\infty \leq X \leq x) \\ &= \int_{-\infty}^x f_X(z) dz. \end{aligned}$$

Just as in the discrete case, the distribution function is defined for all $x \in \mathbb{R}$, even if the sample space S_X is not the whole of the real line.

♦ Properties ♦

- (1) Since it represents a probability, $F_X(x) \in [0, 1]$.
- (2) $F_X(-\infty) = 0$ and $F_X(\infty) = 1$.
- (3) If $a < b$, then $F_X(a) \leq F_X(b)$. ie. $F_X(\cdot)$ is a non-decreasing function.
- (4) When X is continuous, $F_X(x)$ is *continuous*.

Also, by the Fundamental Theorem of Calculus, we have

$$\boxed{\frac{d}{dx} F_X(x) = f_X(x),}$$

and so the *slope* of the CDF $F_X(x)$ is the PDF $f_X(x)$.

Example 6.2

For the light bulb lifetime, X , the distribution function is

$$F_X(x) = \begin{cases} 0, & x < 100 \\ 1 - \frac{100}{x}, & x \geq 100. \end{cases}$$

6.1.4 Median and Quartiles

The *median* of a random quantity is the value m which is the “middle” of the distribution. That is, it is the value m such that

$$P(X \leq m) = P(X \geq m) = \frac{1}{2}.$$

Equivalently, it is the value, m such that

$$F_X(m) = 0.5.$$

Similarly, the *lower quartile* of a random quantity is the value l such that

$$F_X(l) = 0.25,$$

and the *upper quartile* is the value u such that

$$F_X(u) = 0.75.$$

Example 6.3

For the light bulb lifetime, X , what is the median, upper and lower quartile of the distribution?

New Words and Expressions

prescription [pri'skripʃn] *n.* [医]药方，处方；处方药；指示；法规

prescription drugs 处方药

elicit [i'liːt] *vt.* 引出，探出；诱出（回答等）

thereof [ðeər'ɒv] *adv.* 在其中；由此；关于那

vanishingly ['væniʃɪŋli] *adv.* 难以察觉地；趋于零地；消失地

manufacturer [ˌmænjʊ'fæktʃərə(r)] *n.* 制造商，制造厂；厂主；[经]厂商

bulb [bʌlb] *n.* 球茎；电灯泡；[解剖]肿块

Technical Terms

non-decreasing function 非递减函数

upper quartile 上四分位数；上四分位，上四分之一

lower quartile 下四分位数；下四分位，下四分之一

6.2 Properties of Continuous Random Quantities

6.2.1 Expectation and variance of continuous random quantities

The *expectation* or *mean* of a continuous random quantity X is given by

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx,$$

which is just the continuous analogue of the corresponding formula for discrete random quantities.

Similarly, the *variance* is given by

$$\begin{aligned} \text{Var}(X) &= \int_{-\infty}^{\infty} [x - E(X)]^2 f_X(x) dx \\ &= \int_{-\infty}^{\infty} x^2 f_X(x) dx - [E(X)]^2. \end{aligned}$$

Note that the expectation of $g(X)$ is given by

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

and so the variance is just

$$\text{Var}(X) = E([X - E(X)]^2) = E(X^2) - [E(X)]^2$$

as in the discrete case. Note also that all of the properties of expectation and variance derived for discrete random quantities also hold true in the continuous case.

Example 6.4

Consider the random quantity X , with PDF

$$f_X(x) = \begin{cases} \frac{3}{4}(2x - x^2), & 0 < x < 2 \\ 0, & \text{otherwise.} \end{cases}$$

Check that this is a valid PDF (it integrates to 1). Calculate the expectation and variance of X . Evaluate the distribution function. What is the median of this distribution?

6.2.2 PDF and CDF of a Linear Transformation

Let X be a continuous random quantity with PDF $f_X(x)$ and CDF $F_X(x)$, and let $Y = aX + b$ where $a > 0$. What is the PDF and CDF of Y ? It turns out to be easier to work out the CDF first:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(aX + b \leq y) \\ &= P\left(X \leq \frac{y-b}{a}\right) && (\text{since } a > 0) \\ &= F_X\left(\frac{y-b}{a}\right). \end{aligned}$$

So,

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right),$$

and by differentiating both sides with respect to y we get

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Example 6.5

For the light bulb lifetime, X , what is the density of $Y = X/24$, the lifetime of the bulb in days?

New Words and Expressions

differentiate [ˌdɪfə'renʃiət] vt. 表明……间的差别，构成……间差别的特征；[数学] 求……的微分；计算导数或（函数的）微分。现在分词 differentiating；第三人称单数 differentiates

6.3 The Uniform Distribution

Now that we understand the basic properties of continuous random quantities, we can look at some of the important standard continuous probability models. The simplest of these is the uniform distribution.

The random quantity X has a *uniform distribution* over the range $[a, b]$, written

$$X \sim U(a, b)$$

if the PDF is given by

$$f_X(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Thus if $x \in [a, b]$,

$$\begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(y) dy \\ &= \int_{-\infty}^a f_X(y) dy + \int_a^x f_X(y) dy \\ &= 0 + \int_a^x \frac{1}{b-a} dy \\ &= \left[\frac{y}{b-a} \right]_a^x \\ &= \frac{x-a}{b-a}. \end{aligned}$$

Therefore,

$$F_X(x) = \begin{cases} 0, & x < a, \\ \frac{x-a}{b-a}, & a \leq x \leq b, \\ 1, & x > b. \end{cases}$$

We can plot the PDF and CDF in order to see the “shape” of the distribution. Below are plots for $X \sim U(0, 1)$, see Figure 6.1 and Figure 6.2.

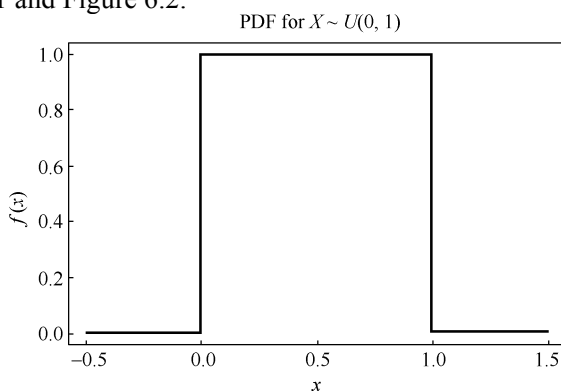


Figure 6.1 PDF $X \sim U(0, 1)$

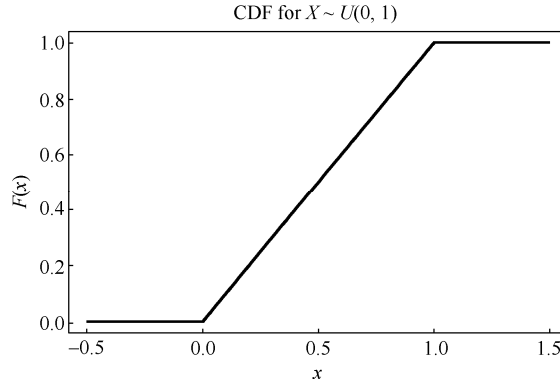


Figure 6.2 CDF $X \sim U(0, 1)$

Clearly the lower quartile, median and upper quartile of the uniform distribution are

$$\frac{3}{4}a + \frac{1}{4}b, \quad \frac{a+b}{2}, \quad \frac{1}{4}a + \frac{3}{4}b,$$

respectively. The expectation of a uniform random quantity is

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^a x f_X(x) dx + \int_a^b x f_X(x) dx + \int_b^{\infty} x f_X(x) dx \\ &= 0 + \int_a^b \frac{x}{b-a} dx + 0 \\ &= \left[\frac{x^2}{2(b-a)} \right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{a+b}{2}. \end{aligned}$$

We can also calculate the variance of X . First we calculate $E(X^2)$ as follows:

$$\begin{aligned} E(X^2) &= \int_a^b \frac{x^2}{b-a} \\ &= \left[\frac{x^3}{3(b-a)} \right]_a^b \\ &= \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{b^2 + ab + a^2}{3}. \end{aligned}$$

Now,

$$\begin{aligned} \text{Var}(X) &= E(X^2) - E(X)^2 \\ &= \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{4} \\ &= \frac{4b^2 + 4ab + 4a^2 - 3b^2 - 6ab - 3a^2}{12} \\ &= \frac{1}{12}[b^2 - 2ab + a^2] \\ &= \frac{(b-a)^2}{12}. \end{aligned}$$

The uniform distribution is rather too simple to realistically model actual experimental data, but is very useful for computer simulation, as random quantities from many different distributions can be obtained from $U(0, 1)$ random quantities.

New Words and Expressions

realistically [ˌriːəˈlɪstɪkli] *adv.* 实际地；现实地；明智地；逼真地

Technical Terms

uniform distribution 均匀分布

6.4 The Exponential Distribution

6.4.1 Definition and Properties

The random variable X has an *exponential distribution* with parameter $\lambda > 0$, written

$$X \sim \text{Exp}(\lambda)$$

if it has PDF

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & \text{otherwise.} \end{cases}$$

The distribution function, $F_X(x)$ is therefore given by

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1 - e^{-\lambda x}, & x \geq 0. \end{cases}$$

The PDF and CDF for an $\text{Exp}(1)$ are shown below Figure 6.3 and Figure 6.4.

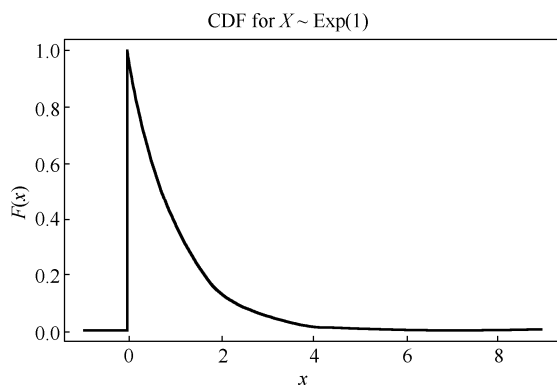


Figure 6.3 PDF $X \sim \text{Exp}(1)$

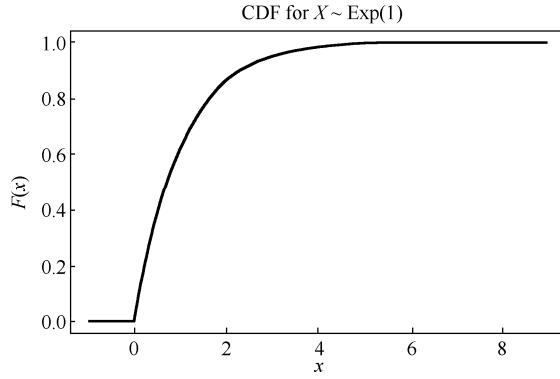


Figure 6.4 CDF $X \sim \text{Exp}(1)$

The expectation of the exponential distribution is

$$\begin{aligned}
 E(X) &= \int_0^{\infty} x\lambda e^{-\lambda x} dx \\
 &= \left[-xe^{-\lambda x} \right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x} dx \quad (\text{by parts}) \\
 &= 0 + \left[\frac{e^{-\lambda x}}{-\lambda} \right]_0^{\infty} \\
 &= \frac{1}{\lambda}.
 \end{aligned}$$

Also,

$$\begin{aligned}
 E(X^2) &= \int_0^{\infty} x^2 \lambda e^{-\lambda x} dx \\
 &= \frac{2}{\lambda^2},
 \end{aligned}$$

and so

$$\begin{aligned}
 \text{Var}(X) &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \\
 &= \frac{1}{\lambda^2}.
 \end{aligned}$$

Note that this means the expectation and standard deviation are both $1/\lambda$.

Notes

- (1) As λ increases, the probability of small values of X increases and the mean decreases.
- (2) The median m is given by

$$m = \frac{\log 2}{\lambda} = \log 2 \, E(X) < E(X).$$

- (3) The exponential distribution is often used to model lifetime and times between random events. The reasons are given below.

6.4.2 Relationship with the Poisson Process

The exponential distribution with parameter λ is the time between events of a Poisson process with rate λ . Let X be the number of events in the interval $(0, t)$. We have seen previously that $X \sim P(\lambda t)$. Let T be the time to the first event. Then

$$\begin{aligned}
F_T(t) &= P(T \leq t) \\
&= 1 - P(T > t) \\
&= 1 - P(X = 0) \\
&= 1 - \frac{(\lambda t)^0 e^{-\lambda t}}{0!} \\
&= 1 - e^{-\lambda t}.
\end{aligned}$$

This is the distribution function of an $\text{Exp}(\lambda)$ random quantity, and so $T \sim \text{Exp}(\lambda)$.

Example 6.6

Consider again the Poisson process for calls arriving at an ISP (Internet Service Provider) at rate 5 per minute. Let T be the time between two consecutive calls. Then we have

$$T \sim \text{Exp}(5)$$

and so $E(T) = \text{SD}(T) = 1/5$.

6.4.3 The Memoryless Property

If $X \sim \text{Exp}(\lambda)$, then

$$\begin{aligned}
P(X > (s+t) | X > t) &= \frac{P([X > (s+t)] \cap [X > t])}{P(X > t)} \\
&= \frac{P(X > (s+t))}{P(X > t)} \\
&= \frac{1 - P(X \leq (s+t))}{1 - P(X \leq t)} \\
&= \frac{1 - F_X(s+t)}{1 - F_X(t)} \\
&= \frac{1 - [1 - e^{-\lambda(s+t)}]}{1 - [1 - e^{-\lambda t}]} \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\
&= e^{-\lambda s} \\
&= 1 - [1 - e^{-\lambda s}] \\
&= 1 - F_X(s) \\
&= 1 - P(X \leq s) \\
&= P(X > s).
\end{aligned}$$

So in the context of lifetimes, the probability of surviving a further time s , having survived time t is the same as the original probability of surviving a time s . This is called the “memoryless” property of the distribution. It is therefore the continuous analogue of the geometric distribution, which also has such a property.

New Words and Expressions

memoryless ['meməri:les] *adj.* 无记忆; 无记忆性; 无记忆性的

Technical Terms

exponential distribution 指数分布

6.5 The Normal Distribution

6.5.1 Definition

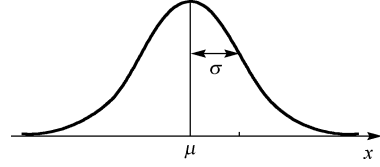
A random quantity X has a normal distribution with parameters μ and σ^2 , written

$$X \sim N(\mu, \sigma^2)$$

if it has probability density function

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\}, \quad -\infty < x < \infty, \quad (6.1)$$

for $\sigma > 0$. Note that $f_X(x)$ is symmetric about $x = \mu$, and so (provided the density integrates to 1), the median of the distribution will be μ . When a graph of all such points is drawn, the normal (bell-shaped) curve will appear as shown in Figure 6.5.



6.5.2 Properties

Checking that the density integrates to one requires the computation of a slightly tricky integral. However, it follows directly from the known “Gaussian” integral

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}}, \quad \alpha > 0,$$

since then

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x) dx &= \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} dx \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp \left\{ -\frac{1}{2\sigma^2} z^2 \right\} dz \quad (\text{putting } z = x - \mu) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \sqrt{\frac{\pi}{1/2\sigma^2}} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \sqrt{2\pi\sigma^2} \\ &= 1. \end{aligned}$$

Now we know that the given PDF represents a valid density, we can calculate the expectation and variance of the normal distribution as follows:

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_{-\infty}^{\infty} x \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2 \right\} dx \\ &= \mu \end{aligned}$$

after a little algebra. Similarly,

$$\begin{aligned}\text{Var}(X) &= \int_{-\infty}^{\infty} (x - \mu)^2 f_X(x) dx \\ &= \int_{-\infty}^{\infty} (x - \mu)^2 \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \mu}{\sigma}\right)^2\right\} dx \\ &= \sigma^2.\end{aligned}$$

Formula (6.2) yields the probability associated with the interval from $x = a$ to $x = b$:

$$P(a \leq x \leq b) = \int_a^b f(x) dx \quad (6.2)$$

The probability that x is within the interval from $x = a$ to $x = b$ is shown as the shaded area in Figure 6.6.

The definite integral of formula (6.2) is a calculus topic and is mathematically more advanced than what is expected in elementary statistics. Instead of using formulas (6.1) and (6.2), we will use a table to find probabilities for normal distributions.

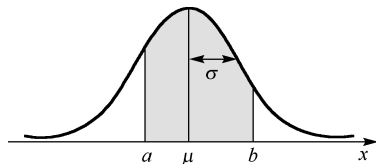


Figure 6.6 Shaded Area: $P(a \leq x \leq b)$

Note: Each different pair of values for the mean, μ , and standard deviation, σ , will result in a different normal probability distribution function.

Formulas (6.1) and (6.2) were used to generate that table. Before we learn to use the table, however, it must be pointed out that the table is expressed in “standardized” form. It is standardized so that this one table can be used to find probabilities for all combinations of mean, μ , and standard deviation, σ , values. That is, the normal probability distribution with mean 38 and standard deviation 7 is similar to the normal probability distribution with mean 123 and standard deviation 32. Recall the empirical rule and the percentages of the distribution that fall within certain intervals of the mean. The same three percentages hold true for all normal distributions.

New Words and Expressions

bell-shaped ['bɛlʃ'eɪpt] *adj.* 钟形的

tricky ['trɪki] *adj.* (形势、工作等) 复杂的; 机警的; 微妙的

standardized ['stændədaɪzd] *adj.* 标准的, 定型的

v. 使合乎规格, 使标准化 (standardize 的过去式和过去分词)

Technical Terms

bell-shaped curve 钟形曲线

shaded area 阴影面积

6.6 The Standard Normal Distribution

There are an unlimited number of normal probability distributions, but fortunately they are all

related to one distribution: the **standard normal distribution**. The standard normal distribution is the normal distribution of the standard variable z (called the “standard score” or “z-score”).

6.6.1 Properties of the Standard Normal Distribution

◇ Properties of the Standard Normal Distribution ◇

- (I) The total area under the standard normal curve is equal to 1.
- (II) The distribution is mound and symmetrical; it extends indefinitely in both directions, approaching but never touching the horizontal axis.
- (III) The distribution has a mean of 0 and a standard deviation of 1.
- (IV) The mean divides the area in half 0.50 on each side.
- (V) Nearly all the area is between $z = -3.00$ and $z = 3.00$.

The **standard normal distribution** is expressed as $N(0, 1)$. The PDF and CDF for a $N(0, 1)$ are shown below Figure 6.7 and Figure 6.8.

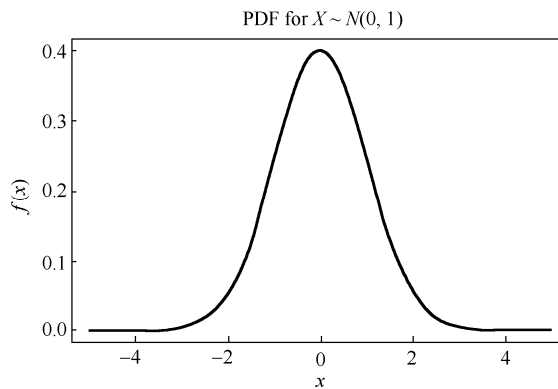


Figure 6.7 PDF for $X \sim N(0, 1)$

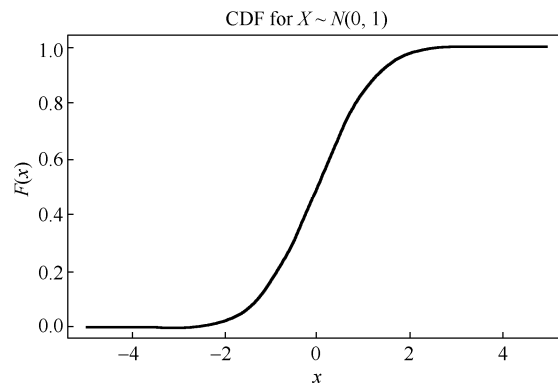


Figure 6.8 CDF for $X \sim N(0, 1)$

Statistical Table 1 in Appendix lists the probabilities associated with the intervals from the mean (located at $z = 0.00$) to a specific value of z . Probabilities of other intervals may be found by using the table entries and the operations of addition and subtraction, in accordance with the

preceding properties. Let's look at several illustrations demonstrating how to use Statistical Table 1 in Appendix to find probabilities of the standard normal score, z .

6.6.2 Finding Area to The Right of $z = 0$

Imagine you want to find the area under the standard normal curve between $z = 0$ and $z = 1.52$ as shown in Figure 6.9.

Statistical Table 1 in Appendix is designed to give the area between $z = 0$ and $z = 1.52$ directly. The z -score is located on the margins, with the units and tenths digits along the left side and the hundredths digit across the top. For $z = 1.52$, locate the row labeled 1.5 and the column labeled 0.02; at their intersection you will find 0.4357, the measure of the area or the probability for the interval $z = 0.00$ to $z = 1.52$ (see Table 6.1 below). Expressed as a probability: $P(0.00 < z < 1.52) = 0.4357$.

Table 6.1 A Portion of Statistical Table 1

z	0.00	0.01	0.02	...
\vdots				
1.5			0.4357	
\vdots				

Recall that one of the basic properties of probability is that the sum of all probabilities is exactly 1.0. Since the area under the normal curve represents the measure of probability, the total area under the bell-shaped curve is exactly 1. This distribution is also symmetrical with respect to the vertical line drawn through $z = 0$, which cuts the area in half at the mean. Can you verify this fact by inspecting formula (6.1)? That is, the area under the curve to the right of the mean is exactly one-half, 0.5, and the area to the left is also one-half, 0.5. Areas (probabilities) not given directly in the table can be found by relying on these facts.

6.6.3 Finding Area in The Right Tail of a Normal Curve

Begin finding the area under the normal curve to the right of $z = 1.52$: $P(z > 1.52)$ by drawing and labeling a sketch. (Sketching and labeling the curve will help you visualize the calculations you are performing). The area to the right of the mean (all of the striped and shaded area in the figure below) is exactly 0.5000. The problem asks for the shaded area that is not included in the 0.4357, see Figure 6.10. Therefore, we subtract 0.4357 from 0.5000:

$$P(z > 1.52) = 0.5000 - 0.4357 = 0.0643$$

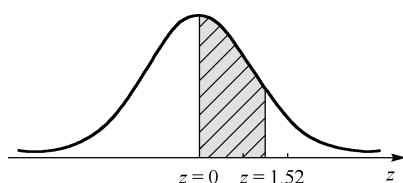


Figure 6.9 Area from $z = 0$ to $z = 1.52$

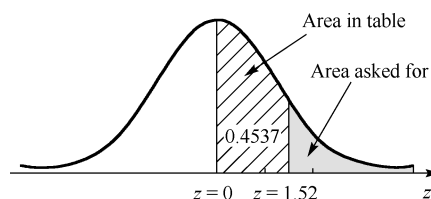


Figure 6.10 The shaded area

Note that we've been writing z with two decimal places and areas and probabilities with four decimal places, as done in Statistical Table 1. Get in the habit of doing so with your work as well.

6.6.4 Finding Area to the Left of a Positive z Value

Let's now examine what's involved in finding the area to the left of $z = 1.52$: $P(z < 1.52)$. The total shaded area is made up of 0.4357 found in the table and the 0.5000 that is to the left of the mean. Therefore, we add 0.4357 to 0.5000, see Figure 6.11:

$$P(z < 1.52) = P(z < 0) + P(0 < z < 1.52) = 0.5000 + 0.4357 = 0.9357$$

Note that the addition and subtraction done here and in the discussion of finding the area in the right tail of a normal curve are correct because the "areas" represent mutually exclusive events.

The symmetry of the normal distribution is a key factor in determining probabilities associated with values below (to the left of) the mean. The area between the mean and $z = -1.52$ is exactly the same as the area between the mean and $z = +1.52$. This fact allows us to find values related to the left side of the distribution, as illustrated in the following examples.

6.6.5 Finding Area from a Negative z to $z = 0$

The area between the mean ($z = 0$) and $z = 2.1$ is the same as the area between $z = 0$ and $z = -2.1$, see Figure 6.12; that is,

$$P(-2.1 < z < 0) = P(0 < z < 2.1)$$

Thus, we have

$$P(-2.1 < z < 0) = P(0 < z < 2.1) = 0.4821$$

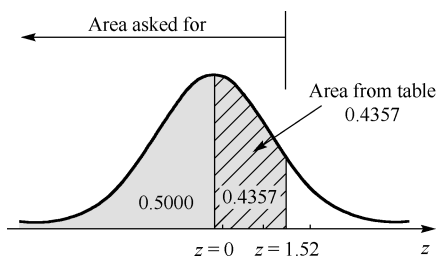


Figure 6.11 The area to the left of $z = 1.52$

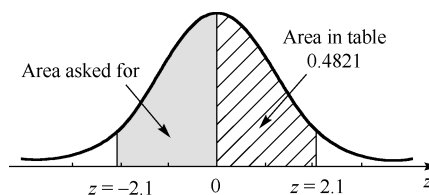


Figure 6.12 Area from a negative z to $z = 0$

6.6.6 Finding Area in the Left Tail of a Normal Curve

The area to the left of $z = -1.35$ is found by subtracting 0.4115 from 0.5000, see Figure 6.13. Therefore, we obtain

$$P(z < -1.35) = P(z < 0) - P(-1.35 < z < 0) = 0.5000 - 0.4115 = 0.0885$$

6.6.7 Finding Area from A Negative z to a Positive z

The area between $z = -1.5$ and $z = 2.1$, $P(-1.5 < z < 2.1)$, is found by adding two areas together,

see Figure 6.14. Both required probabilities are read directly from Statistical Table 1. Therefore, we obtain

$$P(-1.5 < z < 2.1) = P(-1.5 < z < 0) + P(0 < z < 2.1) = 0.4332 + 0.4821 = 0.9153$$

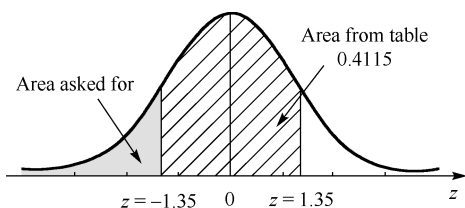


Figure 6.13 The area to the left of $z = -1.35$

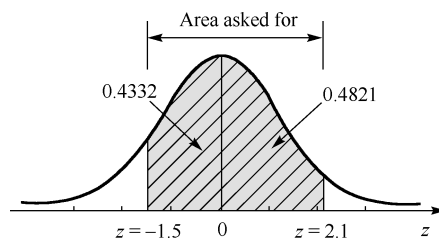


Figure 6.14 Area from a negative z to a positive z

6.6.8 Finding Area Between two z Values of the Same Sign

The area between $z = 0.7$ and $z = 2.1$, $P(0.7 < z < 2.1)$, is found by subtracting. The area between $z = 0$ and $z = 2.1$ includes all the area between $z = 0$ and $z = 0.7$. Therefore, we subtract the area between $z = 0$ and $z = 0.7$ from the area between $z = 0$ and $z = 2.1$, see Figure 6.15.

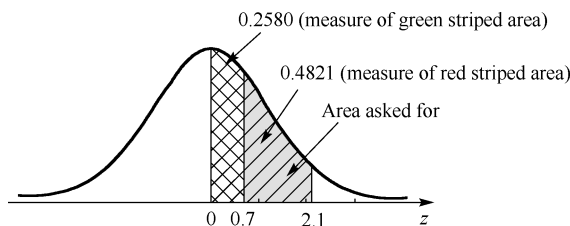


Figure 6.15 Area between two z values of the same sign

Thus, we have

$$P(0.7 < z < 2.1) = P(0 < z < 2.1) - P(0 < z < 0.7) = 0.4821 - 0.2580 = 0.2241$$

The standard normal distribution table can also be used to find a z -score when we are given an area. The next example considers this idea.

6.6.9 Finding z -Scores Associated with a Percentile

Another useful technique used in conjunction with normal distributions is finding the z -score associated with a given percentile. For example, what is the z -score associated with the 75th percentile of a normal distribution? Figure 6.16 gives us that information.

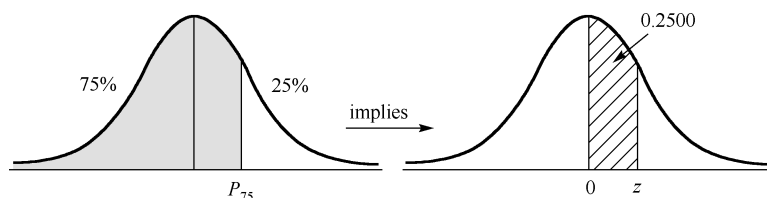


Figure 6.16 P_{75} and Its Associated z -Score

To find this z -score, look in Statistical Table 1 in Appendix and find the “area” entry that is closest to 0.2500; this area entry is 0.2486. Now read the z -score that corresponds to this area.

From the table, the z -score is found to be $z = 0.67$, see Table 6.2. This says that the 75th percentile in a normal distribution is 0.67 (approximately $\frac{2}{3}$) standard deviation above the mean.

Table 6.2 A Portion of Statistical Table 1

z	...	0.07	0	0.08	...
\vdots					
0.6		0.2486	0.2500	0.2517	...
\vdots					

6.6.10 Finding z -scores that Bound an Area

Finding z -scores around partial areas of a normal distribution is also possible. For example, what z -scores bound the middle 95% of a normal distribution? As shown in Figure 6.17, the 95% is split into two equal parts by the mean, so 0.4750 is the area (percentage) between $z = 0$, the mean, and the z -score at the right boundary.

Since we have the area, we look for the entry in Statistical Table 1 closest to 0.4750 (it happens to be exactly 0.4750) and read the z -score, see Table 6.3. We obtain $z = 1.96$.

Table 6.3 A Portion of Statistical Table 1

z	...	0.06	...
\vdots			
1.9		0.4750	...
\vdots			

Therefore, $z = -1.96$ and $z = 1.96$ bound the middle 95% of a normal distribution.

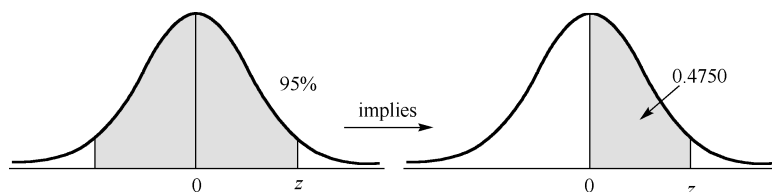


Figure 6.17 Middle 95% of distribution and its associated z -score

New Words and Expressions

touch [tʌtʃ] *vt.* 触摸；使某物与……轻轻接触；吃或喝，尝；[数]与……相切

n. 触摸，碰；触觉，触感；修饰，润色；痕迹

cut [kʌt] *vt.* 削减；剪切；切成；删剪 *n.* 切口；削减；剪裁；切片

boundary ['baʊndri] *n.* 分界线；范围；(球场)边线

Notes

1. 同义词辨析: border, bounds, boundary, frontier, limit 这些名词均含有“边界，边境”之意。

border: 多指国与国之间或两地区的分界处，即分界线附近的边缘部分。

bounds: 常与 boundary 换用，指土地边界，但意思不如 boundary 明确，主要用于抽象事物和文学作品中。

boundary: 侧重地图上正式标定的、双方遵守的边界，也可指较小行政单位间的界线。

frontier: 指两国接壤的前沿地区，属于各国的国境和边疆，多指设防的边界。

limit: 含义广泛，常用作复数，指任何界限、范围、分界线外面的部分，可指有形或无形的东西。

6.7 Applications of Normal Distributions

In section 6.6, we learned how to Statistical Table 1 in Appendix to convert information about the standard normal variable z into probability and vice verse, how to convert probability information about the standard normal distribution into z -scores.

Now we are ready to apply this methodology to all normal distributions. The key is the standard score, z . The information associated with a normal distribution will be in terms of x values or probabilities. We will use the z -score and Statistical Table 1 as the tools to “go between” the given information and the desired answer.

6.7.1 Probabilities and Normal Curves

To demonstrate the process of converting to a standard normal curve to find probabilities, let's consider IQ scores. IQ scores are normally distributed with a mean of 100 and a standard deviation of 16, see Figure 6.18. If a person is picked at random, what is the probability that his or her IQ is between 100 and 115; that is, what is $P(100 < x < 115)$?

Recall that the standard score, z , was defined in Unit 2.

$$\text{In words: } z = \frac{x - (\text{mean of } x)}{\text{standard deviation of } x}$$

$$\text{In algebra: } z = \frac{x - \mu}{\text{standard deviation of } x} \quad (6.3)$$

(Note that when $x = \mu$, the standard score $z = 0$.)

$P(100 < x < 115)$ is represented by the shaded area in the Figure 6.19 below.

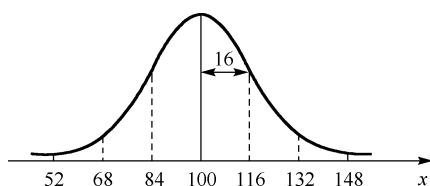


Figure 6.18 The Distribution of IQ

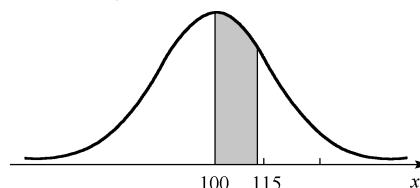


Figure 6.19 The probability of $P(100 < x < 115)$

The variable x must be standardized using formula (6.3). The z values are shown on the next figure.

$$z = \frac{x - \mu}{\sigma}$$

$$\text{when } x = 100 : z = \frac{100 - 100}{16} = 0.00$$

$$\text{when } x = 115 : z = \frac{115 - 100}{16} = 0.94$$

Therefore,

$$P(100 < x < 115) = P(0.00 < z < 0.94) = 0.3264.$$

Thus, the probability is 0.3264 (found by using Statistical Table 1 in Appendix) that a person picked at random has an IQ between 100 and 115, see Figure 6.20.

What if we need to determine a probability for “any” normal curve? How do we calculate probability under “any” normal curve? Let’s continue to use the example of IQ scores and try to find the probability that a person selected at random will have an IQ greater than 90, see Figure 6.21.

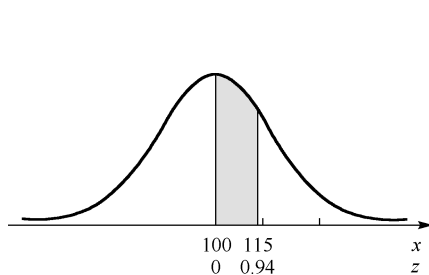


Figure 6.20 The standard score z

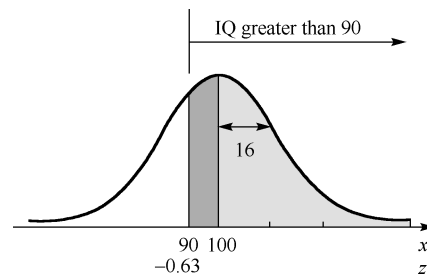


Figure 6.21 The area of an IQ greater than 90

$$z = \frac{x - \mu}{\sigma} = \frac{90 - 100}{16} = \frac{-10}{16} = -0.625 = -0.63$$

$$P(x > 90) = P(z > -0.63) = 0.2357 + 0.5000 = 0.7357$$

Thus, the probability is 0.7357 that a person selected at random will have an IQ greater than 90.

6.7.2 Using the Normal Curve and z

The normal curve is not only an outcome, it can be used with z to determine data values, percentiles, and population parameters.

1. Determine Data Values

In a large class, suppose your instructor tells you that you need to obtain a grade in the top 10% of your class to get an A on a particular exam. From past experience, she is able to estimate that the mean and standard deviation on this exam will be 72 and 13, respectively. What will be the minimum grade needed to obtain an A? (Assume that the grades will have an approximately normal distribution.) We can use the normal curve and z to determine that data value.

Start by converting the 10% to information that is compatible with Statistical Table 1 by subtracting:

$$10\% = 0.1000; 0.5000 - 0.1000 = 0.4000$$

Look in Statistical Table 1 to find the value of z associated with the area entry closest to 0.4000; it is $z = 1.28$, see Figure 6.22. Thus,

$$P(z > 1.28) = 0.10$$

Now find the x value that corresponds to $z = 1.28$ by using formula (6.3):

$$z = \frac{x - \mu}{\sigma} : 1.28 = \frac{x - 72}{13}$$

$$x = 72 + 13 \times 1.28 = 72 + 16.64 = 88.64, \text{ or } 89$$

Thus, if you receive an 89 or higher (the data value), you can expect to be in the top 10% (which means an A).

2. Determine Percentiles

Just as you can use the normal curve and z to find data values, you can also use them to find percentiles. Let's return to the example of IQ scores and find the 33rd percentile for IQ scores when $\mu = 100$ and $\sigma = 16$, see Figure 6.23.

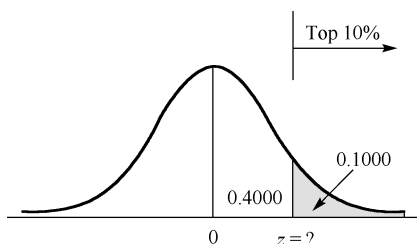


Figure 6.22 IQ greater than 90

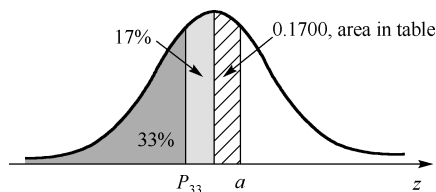


Figure 6.23 The 33rd percentile for IQ scores

$$P(0 < z < \alpha) = 0.17$$

$$\alpha = 0.44 \text{ (cutoff value of } z \text{ from Table 6.4)}$$

$$33\text{rd percentile of } z = -0.44 \text{ (below mean)}$$

Table 6.4 A Portion of Statistical Table 1

z	...	0.04	...
\vdots			
0.4	...	0.1700	...

Now we convert the 33rd percentile of the z -scores, -0.44 , to an x -score using formula (6.3):

$$z = \frac{x - \mu}{\sigma} : -0.44 = \frac{x - 100}{16}$$

$$x - 100 = 16 \times (-0.44)$$

$$x = 100 - 7.04 = 92.96$$

Thus, 92.96 is the 33rd percentile for IQ scores.

3. Determine Population Parameters

The normal curve and z can also be used to determine population parameters. That is, when given related information, you can find the mean, μ . To illustrate how this can be done, let's look at the distribution of junior executives in a large corporation whose incomes are normally distributed with a standard deviation of \$1,200. A cutback is pending, at which time those who earn less than

\$28,000 will be discharged. If such a cut represents 10% of the junior executives, what is the current mean salary of the group of junior executives?

Well, if 10% of the salaries are less than \$28,000, then 40% (or 0.4000) are between \$28,000 and the mean, μ . Statistical Table 1 indicates that $z = -1.28$ is the standard score that occurs at $x = \$28,000$, see Figure 6.24.

Using formula (6.3), we can find the value of μ :

$$z = \frac{x - \mu}{\sigma} : -1.28 = \frac{28000 - \mu}{1200}$$

$$-1536 = 28000 - \mu$$

$$\mu = 28000 + 1536 = \$29536$$

That is, the current mean salary of junior executives is \$29,536.

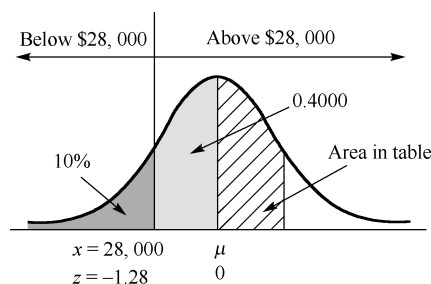


Figure 6.24 10% of the salaries are less than \$28,000

New Words and Expressions

convert [kən'vɜ:t] *vt.* (使) 转变; 使皈依; 兑换, 换算

desired [dɪ'zaɪəd] *adj.* 渴望的, 想得到的

instructor [ɪn'strʌktə(r)] *n.* 指导者, 教师

compatible [kəm'pætəbl] *adj.* 兼容的, 相容的; 和谐的, 协调的 (compatible with 与……不矛盾, 与……相容, 与……一致)

cutback ['kʌtbæk] *n.* 削减, 缩减; (电影等) 倒叙

Technical Terms

cutoff value 临界值, 截断值

junior executive 初级管理人员

6.8 Specific z-score

The z -score is used throughout statistics in a variety of ways; however, the relationship between the numerical value of z and the area under the standard normal distribution curve does not change. Since z will be used with great frequency, we want a convenient notation to identify the necessary information.

The convention that we will use as an “algebraic name” for a specific z -score is $z(\alpha)$, where α represents the “area to the right” of the z being named.

6.8.1 Visual Interpretation of $z(\alpha)$

Figures 6.25 and 6.26 are both visual interpretations of $z(\alpha)$. Figure 6.25 depicts $z(0.05)$ (read “ z of 0.05”), which is the algebraic name for z , such that the area to the right and under the standard normal curve is exactly 0.05. In a similar fashion, Figure 6.26 shows $z(0.60)$ (read “ z of 0.60”), which is the value of z , such that 0.60 of the area lies to its right.

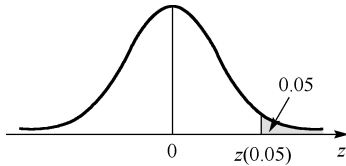


Figure 6.25 Area Associated with $z(0.05)$

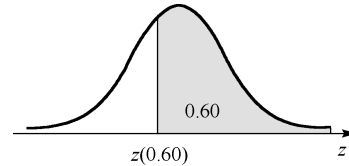


Figure 6.26 Area Associated with $z(0.60)$

6.8.2 Determining Corresponding z Values for $z(\alpha)$

Visual representations like those in Figures 6.25 and 6.26 also have corresponding numerical values. Now let's find the numerical values of $z(0.05)$, $z(0.60)$, and $z(0.95)$.

To find the numerical value of $z(0.05)$, we must convert the area information in the notation into information that we can use with Statistical Table 1 in Appendix. See the areas shown in Figure 6.27. When we look in Statistical Table 1, see Table 6.5, we look for an area as close as possible to 0.4500.

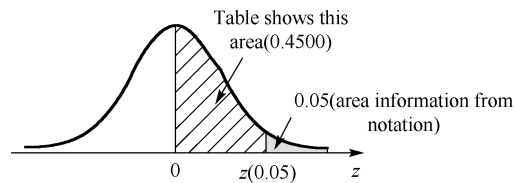


Figure 6.27 Find the value of $z(0.05)$

Table 6.5 A Portion of Statistical Table 1

z	...	0.04		0.05	...
\vdots			\uparrow		
1.6	...	0.4495	0.4500	0.4505	...
\vdots					

Therefore, $z(0.05) = 1.65$.

Note: We will use the z corresponding to the area closest in value. If the value is exactly halfway between the table entries, always use the larger value of z .

To find the numerical value of $z(0.60)$, we must first realize that the value 0.60 is related to Statistical Table 1 by use of the area 0.1000, as shown in Figure 6.28.

The closest values in Statistical Table 1 are 0.0987 and 0.1026.

Therefore, $z(0.60)$ is related to 0.25. Since $z(0.60)$ is below the mean, we conclude that $z(0.60) = -0.25$, see Table 6.6.

Table 6.6 A Portion of Statistical Table 1

z	...	0.05		0.06	...
\vdots			\uparrow		
0.2	...	0.0987	0.1000	0.1026	...
\vdots					

As you would expect, $z(0.95)$ is the diametric opposite of $z(0.05)$, $z(0.95)$ is located on the left-hand side of the normal distribution because the area to the right is 0.95. The area in the tail to the left then contains the other 0.05, as shown in Figure 6.29.

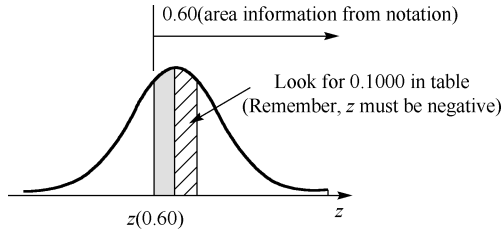


Figure 6.28 Find the value of $z(0.60)$

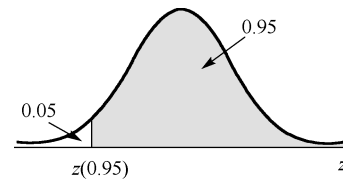


Figure 6.29 Area associated with $z(0.95)$

Because of the symmetrical nature of the normal distribution, $z(0.95)$ is $-z(0.05)$ — that is, $z(0.05)$ with its sign changed. Thus, $z(0.95) = -z(0.05) = 1.65$.

We will use this notation on a regular basis in the following chapters. The values of z that will be used regularly come from one of the following situations: (1) the z -score, such that there is a specified area in one tail of the normal distribution or (2) the z -scores that bound a specified middle proportion of the normal distribution. When the middle proportion of a normal distribution is specified, we can still use the “area to the right” notation to identify the specific z -score involved.

6.8.3 Determining z -scores for Bounded Areas

z -scores can also be determined for bounded areas of a normal distribution. For example, we can find the z -scores that bound the middle 0.95 of the normal distribution. Given 0.95 as the area in the middle (see Figure 6.30), the two tails must contain a total of 0.05. Therefore, each tail contains $1/2$ of 0.05, or 0.025, as shown in Figure 6.31.

In order to find $z(0.025)$ in Statistical Table 1 in Appendix, we must determine the area between the mean and $z(0.025)$. It is $0.5000 - 0.0250 = 0.4750$, as shown in Figure 6.32.

Table 6.7 A Portion of Statistical Table 1

z	...	0.06	...
\vdots			
1.9		0.4750	...
\vdots			

Table 6.7 shows us:

Therefore, $z(0.025) = 1.96$ and $-z(0.975) = z(0.025) = -1.96$. The middle 0.95 of the normal distribution is bounded by -1.96 and 1.96 .

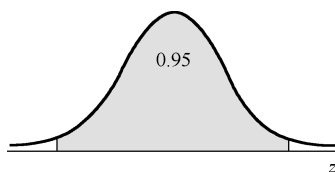


Figure 6.30 Area associated with middle 0.95

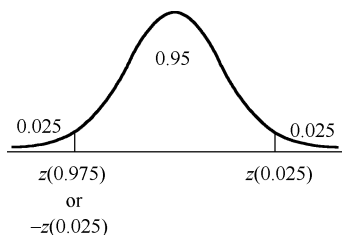


Figure 6.31 Finding z -scores or middle 0.95

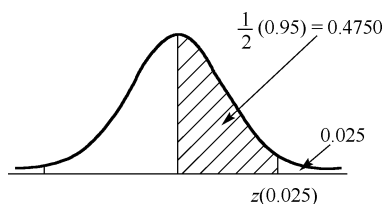


Figure 6.32 Finding the value of $z(0.025)$

New Words and Expressions

convenient [kən'vi:njənt] *adj.* 实用的；便利的；方便的

algebraic [ˌældʒɪ'brenɪk] *adj.* 代数的，代数学的；代数上的

halfway [ˌhɑ:f 'weɪ] *adv.* 在中途；到一半；在中间；大致上

6.9 Normal Approximation of Binomial and Poisson

6.9.1 Normal Approximation of the Binomial

We saw in the last Unit 5 that $X \sim B(n, p)$ could be regarded as the sum of n independent Bernoulli random quantities

$$X = \sum_{k=1}^n I_k,$$

where $I_k \sim \text{Bern}(p)$. Then, because of the central limit theorem, this will be well approximated by a Normal distribution if n is large, and p is not too extreme (if p is very small or very large, a Poisson approximation will be more appropriate). A useful guide is that if

$$0.1 \leq p \leq 0.9 \quad \text{and} \quad n > \max \left[\frac{9(1-p)}{p}, \frac{9p}{1-p} \right]$$

then the binomial distribution may be adequately approximated by a normal distribution. It is important to understand exactly what is meant by this statement. No matter how large n is, the binomial will always be a discrete random quantity with a PMF, whereas the normal is a continuous random quantity with a PDF. These two distributions will always be qualitatively different. The similarity is measured in terms of the CDF, which has a consistent definition for both discrete and continuous random quantities. It is the CDF of the binomial which can be well approximated by a normal CDF. Fortunately, it is the CDF which matters for typical computations involving cumulative probabilities.

When the n and p of a binomial distribution are appropriate for approximation by a normal distribution, the approximation is done by matching expectation and variance. That is

$$B(n, p) \simeq N(np, np[1 - p]).$$

Example 6.8

Reconsider the number of heads X in 100 tosses of an unbiased coin. There $X \sim B(100, 0.5)$, which may be well approximated as

$$X \simeq N(50, 5^2).$$

So, using normal tables we find that $P(40 \leq X \leq 60) \approx 0.955$ and $P(30 \leq X \leq 70) \approx 1.000$, and these are consistent with the exact calculations we undertook earlier: 0.965 and 1.000 respectively.

6.9.2 Normal Approximation of the Poisson

Since the Poisson is derived from the binomial, it is unsurprising that in certain circumstances, the Poisson distribution may also be approximated by the normal. It is generally considered appropriate to make the approximation if the mean of the Poisson is bigger than 20. Again the approximation is done by matching mean and variance:

$$X \sim P(\lambda) \simeq N(\lambda, \lambda) \text{ for } \lambda > 20.$$

Example 6.9

Reconsider the Poisson process for calls arriving at an ISP at rate 5 per minute. Consider the number of calls X , received in 1 hour. We have

$$X \sim P(5 \times 60) = P(300) \simeq N(300, 300).$$

What is the approximate probability that the number of calls is between 280 and 310?

New Words and Expressions

unsurprising [ˌʌnsəˈpraɪzɪŋ] *adj.* 不令人惊讶的；不足为奇的
 approximate [əˈprɒksɪmənt] *v.* 近似，逼近；接近；近似计算
adj. 大约的；近似的；接近的

Technical Terms

approximate probability 近似概率

Problems

6.1 A continuous random variable X , with mean unity, has probability density function $f_x(x)$ given by

$$f_x(x) = \begin{cases} a(b-x)^2, & 0 \leq x \leq b, \\ 0, & \text{otherwise.} \end{cases}$$

Find the values of a and b .

6.2 A continuous random variable X has probability density function $f_x(x)$ given by

$$f_x(x) = \begin{cases} k(2-x)(x-5), & 2 \leq x \leq 5, \\ 0, & \text{elsewhere.} \end{cases}$$

Find the values of k , and hence deduce the mean and variance of X . What are the values of mode and median of the distribution of X ?

6.3 Guaranteed life of a machine

The lifetime in hours of a certain component of a machine has the continuous probability density function

$$f(x) = \frac{1}{1000} e^{-x/1000}, \quad x \geq 0$$

The machine contains five similar components, the lifetime of each having the above distribution. The makers are considering offering a guarantee that not more than two of the original components will have to be replaced during the first 1000 hours of use. Find the probability that such a guarantee would be violated, assuming that the components wear out independently, and that if a component does fail then the replacement used is of particularly high quality and will certainly last for the 1000 hour.

6.4 Find the probability that a data value picked at random from a normal population will have a standard score(z) that lies between the following pairs of z -values.

a. $z = 0$ to $z = 2.10$

b. $z = 0$ to $z = 2.57$

c. $z = 0$ to $z = -1.20$

d. $z = 0$ to $z = -1.57$

6.5 Find the area under the standard normal curve to the left of $z = 1.73$, $P(z < 1.73)$.

6.6 Find the area under the standard normal curve to the left of $z = -1.53$, $P(z < -1.53)$.

6.7 Find the area under the standard normal curve between $z = -2.46$ and $z = 1.46$, $P(-2.46 < z < 1.46)$.

6.8 Find the following:

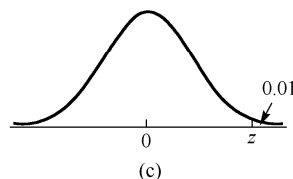
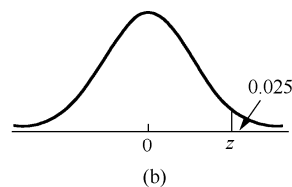
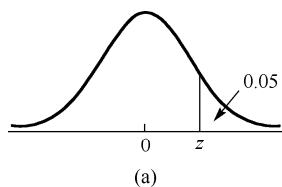
a. $P(0.00 < z < 2.35)$

b. $P(-2.10 < z < 2.34)$

c. $P(z > 0.13)$

d. $P(z < 1.48)$

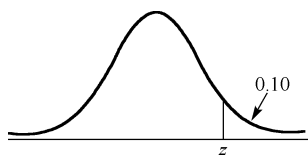
6.9 Find the standard score (z) shown on each of the following diagrams.



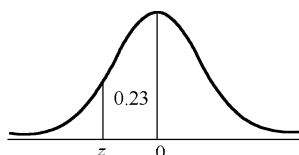
6.10 Find the two z -scores that bound the middle 50% of a normal distribution.

-
- Figure 10.1 consists of five sub-graphs, labeled (a) through (e), each showing a normal distribution curve. In each graph, a vertical line is drawn at a specific z-score on the horizontal axis, and the area under the curve to the right of this line is indicated by an arrow and a numerical value.
- (a) The z-score is 1.5, and the area to the right is 0.03.
 - (b) The z-score is 1.0, and the area to the right is 0.14.
 - (c) The z-score is 0.25, and the area to the right is 0.75.
 - (d) The z-score is 0.5, and the area to the right is 0.98.
 - (e) The z-score is 0.1, and the area to the right is 0.87.

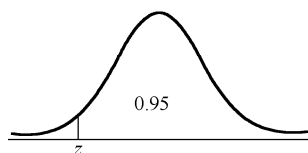
6.19 Using the $z(\alpha)$ notation (identify the value of α used within the parentheses), name each of the standard normal variable z 's shown in the following diagrams.



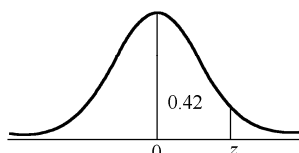
(a)



(b)



(c)



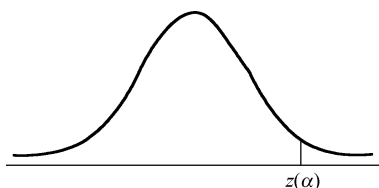
(d)

6.20 Draw a figure of the standard normal curve showing:

a. $z(0.04)$

b. $z(0.94)$

6.21 We are often interested in finding the value of z that bounds a given area in the right-hand tail of the normal distribution, as shown in the accompanying figure. The notation $z(\alpha)$ represents the value of z such that $P(z > z(\alpha)) = \alpha$.



Find the following:

a. $z(0.025)$

b. $z(0.05)$

c. $z(0.01)$

6.22 Use Statistical Table 1 in Appendix to find the following values of z .

a. $z(0.05)$

b. $z(0.01)$

c. $z(0.025)$

d. $z(0.975)$

e. $z(0.98)$

6.23 The z notation, $z(\alpha)$, combines two related concepts — the z -score and the area to the right into a mathematical symbol. Identify the letter in each of the following as being a z -score or being an area; then, with the aid of a diagram, explain what both the given number and the letter represent on the standard normal curve.

a. $z(A) = 0.10$

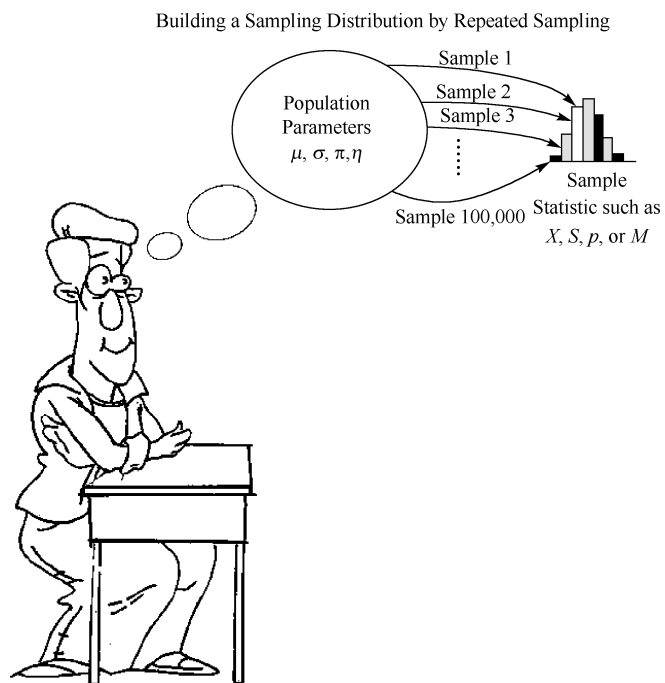
b. $z(0.10) = B$

c. $z(C) = -0.05$

d. $-z(0.05) = D$

Statistics are like a bikini. What they reveal is suggestive, but what they conceal is vital.

—Aaron Levenstein, Professor in Baruch College



Unit 7

Sampling Distributions and CLT



7.1 Sampling Distributions



7.2 The Sampling Distribution of Sample Means



7.3 Application of the Sampling Distribution of Sample Means



7.4 Advanced Central Limit Theorem



Reading English Materials



Problems

7.1 Sampling Distributions

To make inference about a population, we need to discuss sample results little more.

A sample mean \bar{x} is obtained from a sample. Do you expect that this value, \bar{x} , is exactly equal to the value of the population mean μ ? Your answer should be no. We do not expect the means to be identical, but we will be satisfied with our sample results if the sample mean is “close” to the value of the population mean.

Let’s consider a second question: If a second sample is taken, will the second sample have a mean equal to the population mean? Equal to the first sample mean? Again, no, we do not expect the sample mean to be equal to the population mean, nor do we expect the second sample mean to be a repeat of the first one. We do, however, again expect the values to be “close”. This argument should hold for any other sample statistic and its corresponding population value.

The next questions should already have come to mind: What is “close”? How do we determine and measure this closeness? Just how will **repeated sample statistics** be distributed? To answer these questions we must look at *a sampling distribution*. The **sampling distribution of a sample statistic** is the distribution of values for a sample statistic obtained from repeated samples, all of the same size and all drawn from the same population

Let’s start by investigating two different small theoretical sampling distributions. In the first, we’ll introduce a basic sampling distribution of means; in the second, we’ll examine sampling distribution of means in greater detail.

7.1.1 Forming a Sampling Distribution of Means

Definition 1

- **Sampling distribution of a sample statistic:** The distribution of values for a sample statistic obtained from repeated samples, all of the same size and all drawn from the same population.

Example 7.1

Let’s consider a very small, finite population to illustrate the concept of a sampling distribution: the set of single-digit even integers, $\{0, 2, 4, 6, 8\}$, and all possible samples of size 2. We will look at a sampling distribution that might be formed: the sampling distribution of sample means.

First we need to list all possible samples of size 2; there are 25 possible samples, see Table 7.1.

Table 7.1 All possible outcomes of samples of size 2

{0, 0}	{2, 0}	{4, 0}	{6, 0}	{8, 0}
{0, 2}	{2, 2}	{4, 2}	{6, 2}	{8, 2}
{0, 4}	{2, 4}	{4, 4}	{6, 4}	{8, 4}
{0, 6}	{2, 6}	{4, 6}	{6, 6}	{8, 6}
{0, 8}	{2, 8}	{4, 8}	{6, 8}	{8, 8}

Each of these samples has a mean \bar{x} , see Table 7.2. These means are, respectively:

Table 7.2 The mean of each of these outcomes of samples of size 2

0	1	2	3	4
1	2	3	4	5
2	3	4	5	6
3	4	5	6	7
4	5	6	7	8

Each of these samples is equally likely, and thus each of the 25 sample means can be assigned a probability of $1/25 = 0.04$. The sampling distribution of sample means is shown in Table 7.3 as a probability distribution and in Figure 7.1 as a histogram.

Table 7.3 Probability Distribution: Sampling Distribution of Sample Means

X	$P(\bar{x})$
0	0.04
1	0.08
2	0.12
3	0.16
4	0.20
5	0.16
6	0.12
7	0.08
8	0.04

The example above is theoretical in nature and therefore expressed in probabilities. Since this population is small, it is easy to list all 25 possible samples of size 2 (a sample space) and assign probabilities. It is not always possible to do this.

Now, let's empirically (that is, by experimentation) investigate another sampling distribution.

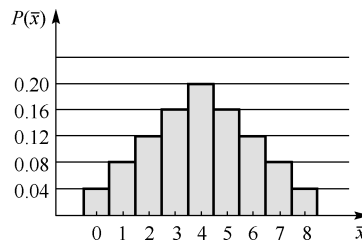


Figure 7.1 Histogram: Sampling Distribution of Sample Means

7.1.2 Creating a Sampling Distribution of Sample Means

Example 7.2

Let's consider a population that consists of five equally likely integers: 1, 2, 3, 4, and 5. We can observe a portion of the sampling distribution of sample means when 30 samples of size 5 are randomly selected. Figure 7.2 shows a histogram representation of the population.

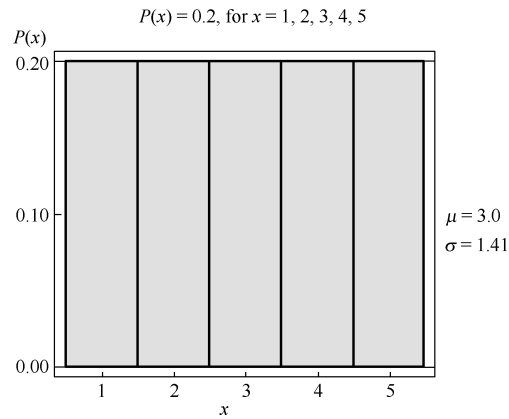


Figure 7.2 The Population: Theoretical Probability Distribution

Table 7.4 shows 30 samples and their means. The resulting sampling distribution, a **frequency distribution**, of sample means is shown in Figure 7.3. Notice that this distribution of sample means does not look like the population. Rather, it seems to display the characteristics of a normal distribution; it is mounded and nearly symmetric about its mean (approximately 3.0).

Table 7.4 30 Samples of Size 5

No.	Sample	\bar{x}	No.	Sample	\bar{x}
1	4, 5, 1, 4, 5	3.8	16	4, 5, 5, 3, 5	4.4
2	1, 1, 3, 5, 1	2.2	17	3, 3, 1, 2, 1	2.0
3	2, 5, 1, 5, 1	2.8	18	2, 1, 3, 2, 2	2.0
4	4, 3, 3, 1, 1	2.4	19	4, 3, 4, 2, 1	2.8
5	1, 2, 5, 2, 4	2.8	20	5, 3, 1, 4, 2	3.0
6	4, 2, 2, 5, 4	3.4	21	4, 4, 2, 2, 5	3.4
7	1, 4, 5, 5, 2	3.4	22	3, 3, 5, 3, 5	3.8
8	4, 5, 3, 1, 2	3.0	23	3, 4, 4, 2, 2	3.0
9	5, 3, 3, 3, 5	3.8	24	3, 3, 4, 5, 3	3.6
10	5, 2, 1, 1, 2	2.2	25	5, 1, 5, 2, 3	3.2
11	2, 1, 4, 1, 3	2.2	26	3, 3, 3, 5, 2	3.2
12	5, 4, 3, 1, 1	2.8	27	3, 4, 4, 4, 4	3.8
13	1, 3, 1, 5, 5	3.0	28	2, 3, 2, 4, 1	2.4
14	3, 4, 5, 1, 1	2.8	29	2, 1, 1, 2, 4	2.0
15	3, 1, 5, 3, 1	2.6	30	5, 3, 3, 2, 5	3.6

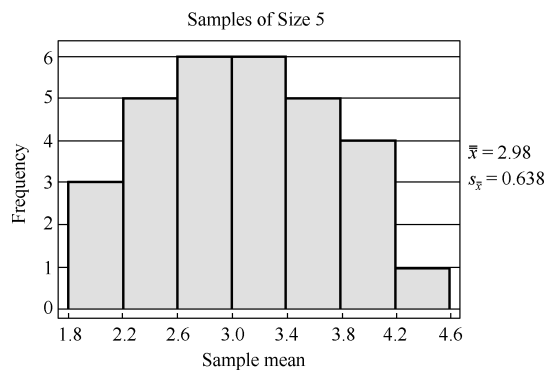


Figure 7.3 Empirical Distribution of Sample Means

Note: The variable for the sampling distribution is \bar{x} ; therefore, the mean of the \bar{x} 's is $\bar{\bar{x}}$ and the standard deviation of \bar{x} is $s_{\bar{x}}$.

The theory involved with sampling distributions that will be described in the remainder of this unit requires *random sampling*. Recall from Unit 1 that a *random sample* is obtained in such a way that each possible sample of fixed size n has an equal probability of being selected.

Figure 7.4 shows how the sampling distribution of sample means is formed.

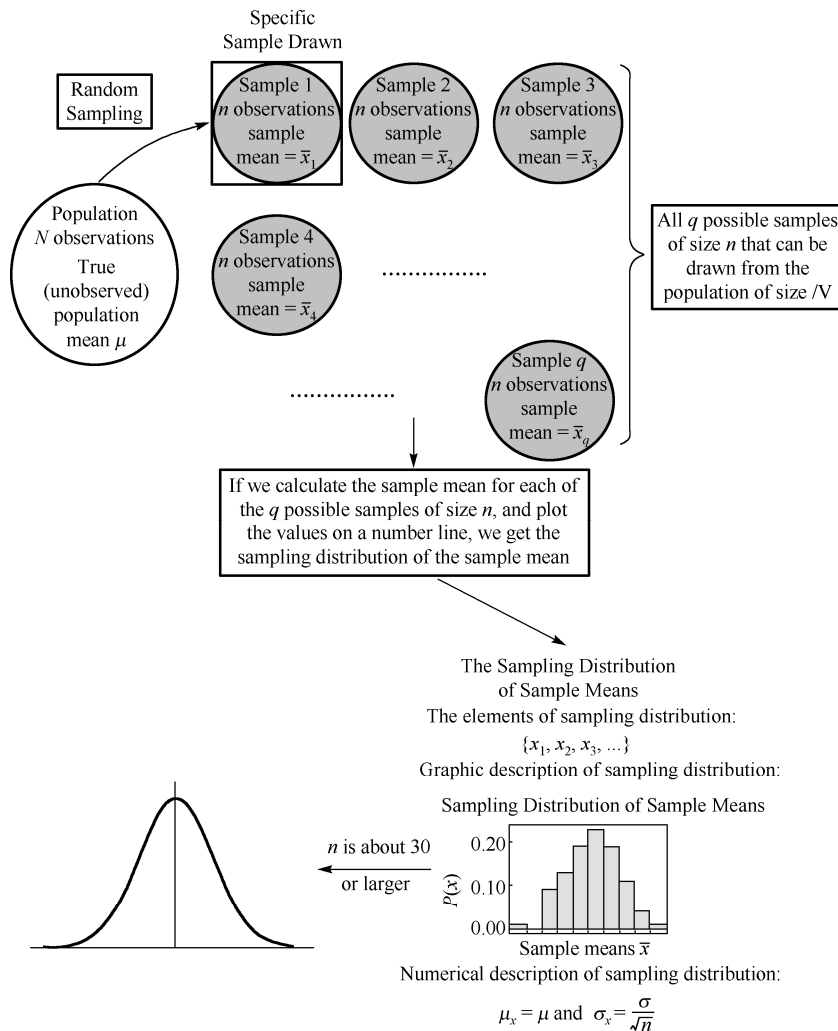


Figure 7.4 The sample distribution of sample means

Example 7.3 Average Age of Urban Transit Rail Vehicles

There are many reasons for collecting data repeatedly. Not all repeated data collections are performed to form a sampling distribution. Consider the “Average Age of Urban Transit Rail Vehicles (Years)” statistics from the U.S. Department of Transportation that follow. The Table 7.5 shows the average age for four different classifications of transit rail vehicles tracked over several years. By studying the pattern of change in the average age for each class of vehicle, a person can

draw conclusions about what has been happening to the fleet over several years. Chances are the people involved in maintaining each fleet can also detect when a change in policies regarding replacement of older vehicles is needed. However useful this information is, there is no sampling distribution involved here.

Table 7.5 Average Age of Urban Transit Rail Vehicles (Years)

	1985	1990	1995	2000	2003
Transit rail					
Commuter rail locomotives ^a	16.3	15.7	15.9	13.4	16.6
Commuter rail passenger coaches	19.1	17.6	21.4	16.9	20..5
Heavy-rail passenger cars	17.1	16.2	19.3	22.9	19.0
Light-rail vehicles (streetcars)	20.6	15.2	16.8	16.1	15.6

^a Locomotives used in Amtrak intercity passenger services are not included.

New Words and Expressions

single-digit ['sɪŋɡld'ɪdʒɪt] 单位数，个位数，一位数

little more 多一点，更多一点，多一些

vehicle ['vi:əkl] *n.* 车辆；交通工具；传播媒介，媒介物

coach [kəʊtʃ] *n.* 教练；(铁路)旅客车厢；长途客运汽车；四轮大马车

track [træk] *vt.* 跟踪；监看，监测；追踪 *vi.* 沿着轨道前进；沿着一条路走，旅行

streetcar ['stri:tka:(r)] *n.* 有轨电车

fleet [fli:t] *n.* 船队；车队；港湾，小河

Technical Terms

sampling distribution 抽样分布

random sample 随机样本

Notes

U.S. Department of Transportation 美国交通部

Urban Transit 城市运输，城市公共交通

7.2 The Sampling Distribution of Sample Means

On the preceding pages we discussed the sampling distributions of sample means, and many other sampling distributions could be discussed.

The only one of concern to us at this time, however, is the **sampling distribution of sample means (SDSM)**:

If all possible random samples, each of size n , are taken from any population with mean μ and standard deviation σ , then the sampling distribution of sample means will have the following:

(i) A mean $\mu_{\bar{x}}$ equal to μ ,

(ii) A standard deviation $\sigma_{\bar{x}}$ equal to $\frac{\sigma}{\sqrt{n}}$.

Furthermore, if the sampled population has a normal distribution, then the sampling distribution of \bar{x} will also be normal for samples of all sizes.

The two-part statement in the box above is very interesting. The first part tells us about the relationship between the population mean and standard deviation, and the sampling distribution mean and standard deviation for all sampling distributions of sample means. The standard deviation of the sampling distribution is denoted by $\sigma_{\bar{x}}$ and given a specific name to avoid confusion with the population standard deviation, σ . We call $\sigma_{\bar{x}}$ the standard error of the mean.

Definition 2

■ **Standard error of the mean ($\sigma_{\bar{x}}$):** The standard deviation of the sampling distribution of sample means.

◇ Central limit theorem (CLT) ◇

The sampling distribution of sample means will more closely resemble the normal distribution as the sample size increases.

The second part indicates that this information is not always useful. Stated differently, it says that the mean value of only a few observations will be normally distributed when samples are drawn from a normally distributed population but will not be normally distributed when the sampled population is uniform, skewed, or otherwise not normal. However, the central limit theorem gives us some additional and very important information about the sampling distribution of sample means. According to the **central limit theorem (CLT)**, the sampling distribution of sample means will more closely resemble the normal distribution as the sample size increases.

If the sampled distribution is normal, then the sampling distribution of sample means (SDSM) is normal, as stated above, and the central limit theorem (CLT) does not apply. But if the sampled population is not normal, the sampling distribution will still be approximately normally distributed under the right conditions. If the sampled distribution is nearly normal, the \bar{x} distribution is approximately normal for fairly small n (possibly as small as 15). When the sampled distribution lacks symmetry, n may have to be quite large (maybe 50 or more) before the normal distribution provides a satisfactory approximation.

7.2.1 Central Limit Theorem

Abraham de Moivre was a pioneer in the theory of probability and published *The Doctrine of Chance* in Latin in 1711 and then in expanded editions later in the century. The 1756 edition

contained his most important contribution the approximation of the binomial distributions for a large number of trials using the normal distribution. The definition of statistical independence also made its debut along with many dice and other games, de Moivre proved that the central limit theorem holds for numbers resulting from games of chance. With the use of mathematics, he also successfully predicted the date of his own death.

By combining the preceding information, we can describe the sampling distribution of \bar{x} completely: (1) the location of the center (mean), (2) a measure of spread indicating how widely the distribution is dispersed (standard deviation), and (3) an indication of how it is distributed.

(i) $\mu_{\bar{x}} = \mu$; the mean of the sampling distribution ($\mu_{\bar{x}}$) is equal to the mean of the population (μ).

(ii) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$; the standard error of the mean ($\sigma_{\bar{x}}$) is equal to the standard deviation of the population (σ) divided by the square root of the sample size, n .

(iii) The distribution of sample means is normal when the parent population is normally distributed, and the CLT tells us that the distribution of sample means becomes approximately normal (regardless of the shape of the parent population) when the sample size is large enough.

Note: The n referred to is the size of each sample in the sampling distribution. (The number of repeated samples used in an empirical situation has no effect on the standard error.)

We do not show the proof for the preceding three facts in this text; however, their validity will be demonstrated by examining two illustrations. For the first illustration, let's consider a population for which we can construct the theoretical sampling distribution of all possible samples.

7.2.2 Constructing a Sampling Distribution of Sample Means

Example 7.4

Let's consider all possible samples of size 2 that could be drawn from a population that contains the three numbers 2, 4, and 6. First let's look at the population itself: Construct a histogram to picture its distribution, see Figure 7.5; calculate the mean μ and the standard deviation σ , see Table 7.6. Remember: We must use the techniques from Unit 5 for discrete probability distributions.

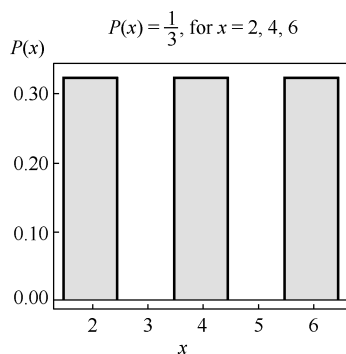


Figure 7.5 Population

Table 7.6 Extensions Table for x

x	$P(x)$	$xP(x)$	$x^2P(x)$
2	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{4}{3}$
4	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{16}{3}$
6	$\frac{1}{3}$	$\frac{6}{3}$	$\frac{36}{3}$
	$\frac{3}{3}$	$\frac{12}{3}$	$\frac{56}{3}$
	1.0	4.0	18.66

$$\mu = 4.0$$

$$\sigma = \sqrt{18.6\bar{6} - (4.0)^2} = \sqrt{2.6\bar{6}} = 1.63$$

Table 7.7 lists all the possible samples of size 2 that can be drawn from this population. One number is drawn, observed, and then returned to the population before the second number is drawn. Table 7.7 also lists the means of these samples. The probability distribution for these means and the extensions are given in Table 7.8, along with the calculation of the mean and the standard error of the mean for the sampling distribution. The histogram for the sampling distribution of sample means is shown in Figure 7.6.

Table 7.7 All Nine Possible Samples of Size 2

Sample	\bar{x}	Sample	\bar{x}	Sample	\bar{x}
2, 2	2	4, 2	3	6, 2	4
2, 4	3	4, 4	4	6, 4	5
2, 6	4	4, 6	5	6, 6	6

Table 7.8 Extensions Table for \bar{x}

\bar{x}	$P(\bar{x})$	$\bar{x}P(\bar{x})$	$\bar{x}^2 P(\bar{x})$
2	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{4}{9}$
3	$\frac{2}{9}$	$\frac{6}{9}$	$\frac{18}{9}$
4	$\frac{3}{9}$	$\frac{12}{9}$	$\frac{48}{9}$
5	$\frac{2}{9}$	$\frac{10}{9}$	$\frac{50}{9}$
6	$\frac{1}{9}$	$\frac{6}{9}$	$\frac{36}{9}$
	$\frac{9}{9}$	$\frac{36}{9}$	$\frac{156}{9}$
	1.0	4.0	17.3 $\bar{3}$

$$\mu_{\bar{x}} = 4.0$$

$$\sigma_{\bar{x}} = \sqrt{17.3\bar{3} - (4.0)^2} = \sqrt{1.3\bar{3}} = 1.15$$

Let's now check the truth of the three facts about the sampling distribution of sample means:

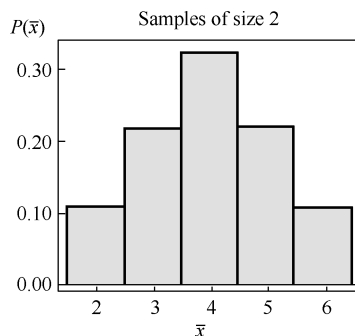


Figure 7.6 Sampling Distribution of Sample Means

Three Facts

(I) The mean $\mu_{\bar{x}}$ of the sampling distribution will equal the mean μ of the population: Both μ and $\mu_{\bar{x}}$ have the value 4.0.

(II) The standard error of the mean $\sigma_{\bar{x}}$ for the sampling distribution will, equal the standard deviation σ of the population divided by the square root of the sample sized, n : $\sigma_{\bar{x}} = 1.15$ and $\sigma = 1.63$, $n = 2$, $\frac{\sigma}{\sqrt{n}} = \frac{1.63}{\sqrt{2}} = 1.15$; they are equal: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

(III) The distribution will become approximately normally distributed: The histogram in Figure 7.6 very strongly suggests normality.

Our example uses a theoretical situation and suggests that all three facts appear to hold true. Do these three facts hold when actual data are collected? Let's look back at the example using five equally likely integers (1, 2, 3, 4, 5) from Section 7.1 and see if all three facts are supported by the empirical sampling distribution there.

First, let's look at the population in the theoretical probability distribution from which the samples were taken. Figure 7.2 is a histogram showing the probability distribution for randomly selected data from the population of equally likely integers 1, 2, 3, 4, 5. The population mean μ equals 3.0. The population standard deviation σ is $\sqrt{2}$, or 1.41. The population has a uniform distribution.

Now let's look at the empirical distribution of the 30 sample means found in our earlier example. From the 30 values of \bar{x} in Table 7.2, the observed mean of the \bar{x} 's, $\bar{\bar{x}}$, is 2.98 and the observed standard error of the mean, $s_{\bar{x}}$, is 0.638. The histogram of the sampling distribution in Figure 7.3 appears to be mound, approximately symmetrical, and centered near the value 3.0.

Now let's check the truth of the three specific properties:

(i) $\mu_{\bar{x}}$ and μ will be equal: The mean of the population μ is 3.0, and the observed sampling distribution mean $\bar{\bar{x}}$ is 2.98; they are very close in value.

(ii) $\sigma_{\bar{x}}$ will equal $\frac{\sigma}{\sqrt{n}}$. $\sigma = 1.41$ and $n = 5$; therefore, $\frac{\sigma}{\sqrt{n}} = \frac{1.41}{\sqrt{5}} = 0.632$, and $s_{\bar{x}} = 0.638$; they are very close in value. (Remember that we have taken only 30 samples, not all possible samples, of size 5.)

(iii) The sampling distribution of \bar{x} will be approximately normally distributed. Even though the population has a rectangular distribution, the histogram in Figure 7.3 suggests that the \bar{x} distribution has some of the properties of normality (mound, symmetric).

Although our examples do not constitute a proof, the evidence seems to strongly suggest that both statements, the sampling distribution of sample means and the central limit theorem, are true.

Having taken a look at these two specific illustrations, let's now look at four graphic illustrations that present the sampling distribution information and the CLT in a slightly different form. Each of these illustrations has four distributions. The first graph shows the distribution of the parent population, the distribution of the individual x values. Each of the other three graphs shows a sampling distribution of sample means, \bar{x} 's, using three different sample sizes.

In Figure 7.7 we have a uniform distribution, much like that in Figure 7.2 for the integer illustration, and the resulting distributions of sample means for samples of sizes 2, 5, and 30.

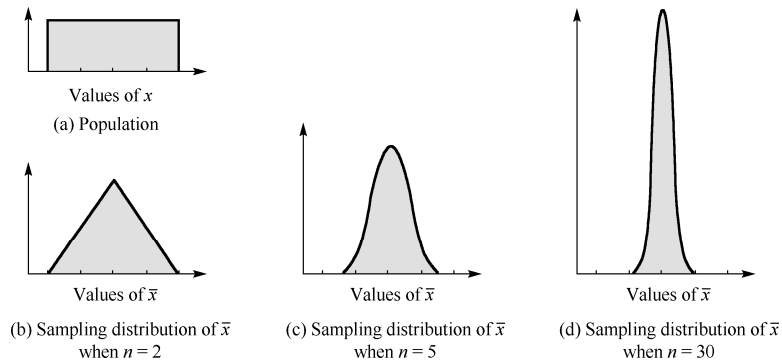


Figure 7.7 Uniform Distribution

Figure 7.8 shows a U-shaped population and the three sampling distributions.

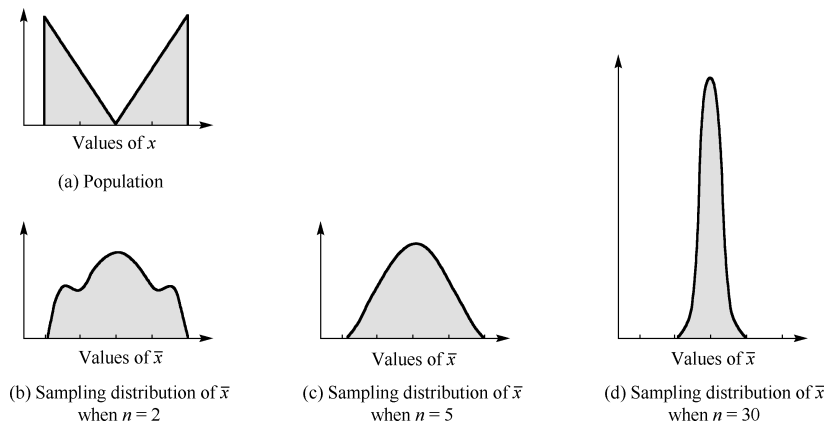


Figure 7.8 U-Shaped Distribution

Figure 7.9 shows a J-shaped population and the three sampling distributions.

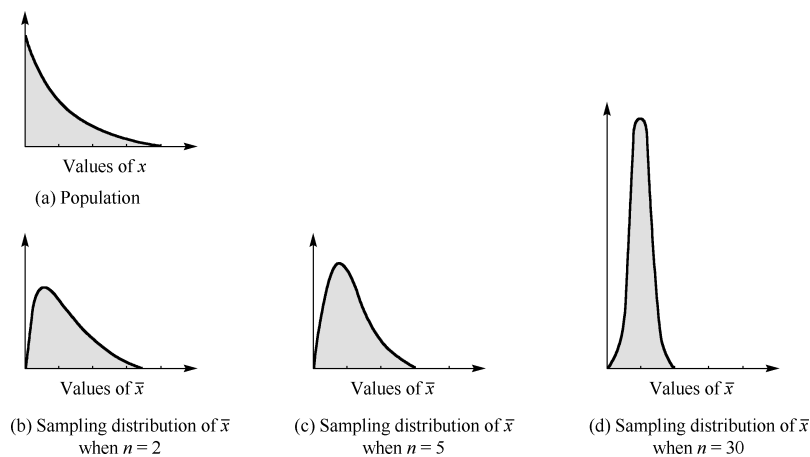


Figure 7.9 J-Shaped Distribution

All three non-normal distributions seem to verify the CLT; the sampling distributions of sample means appear to be approximately normal for all three when samples of size 30 were used. With the normal population, see Figure 7.10, the sampling distributions for all sample sizes appear to be normal. Thus, you have seen an amazing phenomenon: No matter what the shape of a population, the sampling distribution of sample means either is normal or becomes approximately normal when n becomes sufficiently large.

You should notice one other point: The sample mean becomes less variable as the sample size increases. Notice that as n increases from 2 to 30, all the distributions become narrower.

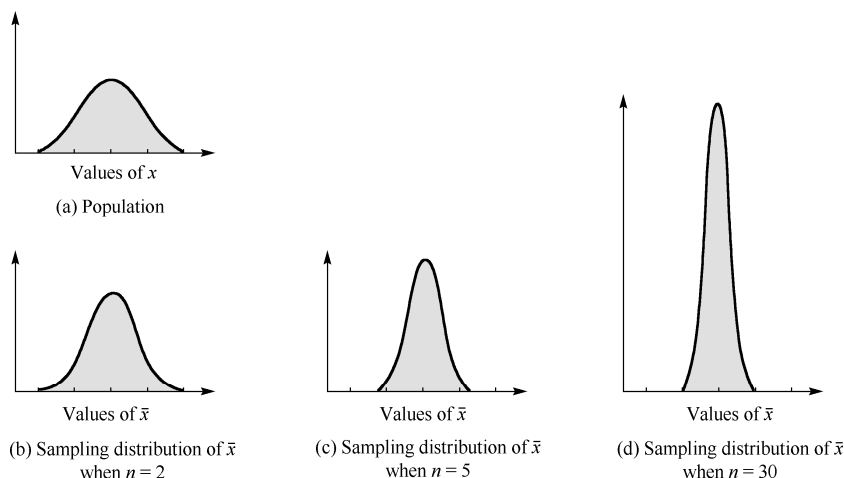


Figure 7.10 Normal Distribution

New Words and Expressions

confusion [kən'fju:ʒn] *n.* 混乱；混淆；困惑

debut ['derbjʊ:] *n.* 初次露面，初次表演，首次出场，处女秀

proof [pru:f] *n.* 证明；校样；检验

Technical Terms

standard error of mean 均值的标准误差

Central Limit Theorem (CLT) 中心极限定理

Notes

1. Abraham de Moivre 亚伯拉罕·棣莫弗 (1667.5.26—1754.11.27), 法国裔英国籍的数学家, 发现了棣莫弗公式, 将复数和三角学联系起来。正态分布最早是棣莫弗在 1711 年出版的著作《机会论》(Doctrin of Change) 及 1734 年发表的一篇关于二项分布文章中提出的, 当二项随机变数的位置参数 n 很大, 并且形状参数 p 为 $1/2$ 时, 则所推导出二项分布的近似分布函数就是正态分布。

拉普拉斯在 1812 年发表的《分析概率论》(Theorie Analytique des Probabilites) 中将棣莫弗的结论扩展到二项分布的位置参数为 n 且形状参数为 $1 > p > 0$ 时。现在, 这一结论通常被称为棣莫佛-拉普拉斯定理。

7.3 Application of the Sampling Distribution of Sample Means

When the sampling distribution of sample means is normally distributed, or approximately normally distributed, we will be able to answer probability questions with the aid of the standard normal distribution (Statistical Table 1 Appendix).

7.3.1 Converting \bar{x} Information into z - scores

When the population is normally distributed, the sampling distribution of \bar{x} 's is normally distributed. To determine probabilities associated with a normal distribution, we will need to format a probability statement involving the z -score in order to use Statistical Table 1 Appendix, the standard normal distribution table. Consider a normal population with $\mu = 100$ and $\sigma = 20$. If a random sample of size 16 is selected, what is the probability that this sample will have a mean value between 90 and 110? That is, what is $P(90 < \bar{x} < 110)$?

This population is normally distributed, so the sampling distribution of \bar{x} 's is normally distributed. We will need to convert the statement $P(90 < \bar{x} < 110)$ to a probability statement involving the z -score. The sampling distribution is shown in the figure, where the shaded area represents $P(90 < \bar{x} < 110)$.

The formula for finding the z -score corresponding to a known value of \bar{x} is

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \quad (7.1)$$

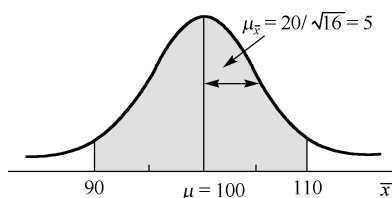


Figure 7.11 A normal population with $\mu = 100$ and $\sigma = 20$

The mean and standard error of the mean are $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Therefore, we will rewrite formula (7.1) in terms of μ , σ , and n :

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad (7.2)$$

Returning to the illustration and applying formula (7.2), we find:

$$z\text{-score for } \bar{x} = 90: z = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{90 - 100}{20 / \sqrt{16}} = \frac{-10}{5} = -2.00$$

$$z\text{-score for } \bar{x} = 110: z = \frac{x - \mu}{\sigma / \sqrt{n}} = \frac{110 - 100}{20 / \sqrt{16}} = \frac{10}{5} = 2.00$$

Therefore,

$$P(90 < \bar{x} < 100) = P(-2.00 < z < 2.00) = 2(0.4772) = 0.9544$$

7.3.2 Distribution of \bar{x} and Increasing Individual Sample Size

Before we look at more illustrations, let's consider what is implied by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. To demonstrate, let's suppose that $\sigma = 20$ and let's use a sampling distribution of samples of size 4. Now $\sigma_{\bar{x}}$ is $20 / \sqrt{4}$, or 10, and approximately 95% (0.9544) of all such sample means should be within the interval from 20 below to 20 above the population mean (within two standard deviations of the population mean). However, if the sample size is increased to 16, $\sigma_{\bar{x}}$ becomes $20 / \sqrt{16} = 5$ and approximately 95% of the sampling distribution should be within 10 units of the mean, and so on. As the sample size increases, the size of $\sigma_{\bar{x}}$ becomes smaller so that the distribution of sample means becomes much narrower. Figure 7.12 illustrates what happens to the distribution of \bar{x} 's as the size of the individual samples increases.

Recall that the area (probability) under the normal curve is always exactly one. So as the width of the curve narrows, the height has to increase in order to maintain this area.

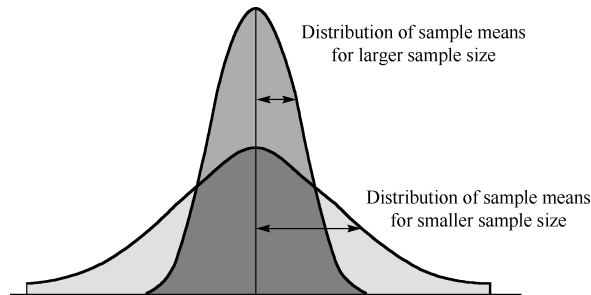


Figure 7.12 Distributions of sample means

Example 7.5 Calculating Probabilities for the Mean

Calculating probabilities is one way we are able to make predictions about the corresponding population parameter we are looking at. Let's use the example of mean heights of kindergarteners to demonstrate. Kindergarten children have heights that are approximately normally distributed about a mean of 39 inches and a standard deviation of 2 inches. A random sample of size 25 is taken and the mean \bar{x} is calculated. What is the probability that this mean value will be between 38.5 and 40.0 in, see Figure 7.13?

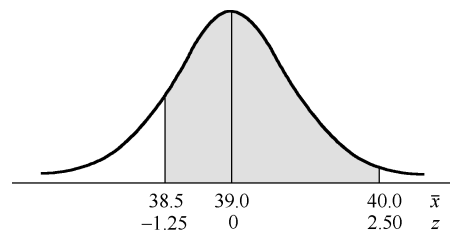


Figure 7.13 Probability of the mean in between 38.5 and 40.0

To find out, we first need to find $P(38.5 < \bar{x} < 40.0)$. The values of \bar{x} , 38.5 and 40.0, must be converted to z -scores (necessary for use of Table 1) using $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$:

$$\bar{x} = 38.5 : \quad z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{38.5 - 39.0}{2 / \sqrt{25}} = \frac{-0.5}{0.4} = -1.25$$

$$\bar{x} = 40.0 : \quad z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{40.0 - 39.0}{2 / \sqrt{25}} = \frac{1.0}{0.4} = 2.50$$

Therefore,

$$P(38.5 < \bar{x} < 40.0) = P(-1.25 < z < 2.50) = 0.3944 + 0.4938 = 0.88 / 82$$

In the same vein, we can calculate mean height limits for a certain portion of our population. Using the heights of kindergarten children given in the previous example, we can figure out limits within which the middle 90% of the sampling distribution of sample means for samples of size 100 falls.

The two tools we have to work with are formula (7.2) and Statistical Table 1. The formula relates the key values of the population to the key values of the sampling distribution, and Statistical Table 1 relates areas to z -scores. First, using Statistical Table 1, see Table 7.9, we find that the middle 0.9000 is bounded by $z = \pm 1.65$.

Table 7.9 A Portion of Statistical Table 1

z	...	0.04		0.05	...
\vdots					
1.6	...	0.4495	0.4500	0.4505	...
\vdots					

Second, we use formula (7.2), $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$:

$$\begin{aligned} z = -1.65 : \quad -1.65 &= \frac{\bar{x} - 39.0}{2 / \sqrt{100}} \\ \bar{x} - 39 &= (-1.65)(0.2) \\ \bar{x} &= 39 - 0.33 \\ &= 38.67 \\ z = 1.65 : \quad 1.65 &= \frac{\bar{x} - 39.0}{2 / \sqrt{100}} \\ \bar{x} - 39 &= (1.65)(0.2) \\ \bar{x} &= 39 + 0.33 \\ &= 39.33 \end{aligned}$$

Thus,

$$P(38.67 < \bar{x} < 39.33) = 0.90$$

Therefore, 38.67 in. and 39.33 in. are the limits that capture the middle 90% of the sample means.

The basic purpose for considering what happens when a population is repeatedly sampled, as discussed in this unit, is to form sampling distributions. We will begin to make inferences about population means and the values of population parameters in Unit 8.

New Words and Expressions

with the aid of 借助于，通过.....的帮助

vein [veɪn] *n.* 静脉；[地]矿脉，岩脉；气质，倾向

in the same vein 以同样的方式；以同样的风格

kindergartener ['kɪndəɡɑ:tənər] *n.* 幼儿园教师，幼儿园里的小孩

kindergarten ['kɪndəɡɑ:tn] *n.* 幼儿园，幼稚园，学前班。in kindergarten 在幼儿园，
kindergarten education 幼教，kindergarten children 幼儿园孩子

figure out 计算出；合计；解决

7.4 Advanced Central Limit Theorem

Generally, regardless of the population distribution model, as the sample size increases, the sample mean tends to be normally distributed around the population mean, and its standard deviation shrinks as n increases.

The Central Limit Theorem provides us with a shortcut to the information required for constructing a sampling distribution. By applying the CLT we can obtain the descriptive values for a sampling distribution (usually, the mean and the standard error, which is computed from the sampling variance) and we can also obtain probabilities associated with any of the sample means in the sampling distribution.

In fact, certain conditions must be met to use the CLT. That is (i) the samples must be independent; (ii) the sample size must be “big enough”.

◇ CLT conditions ◇

(i) Independent Samples Test:

◆ “Randomization”: Each sample should represent a random sample from the population, or at least follow the population distribution.

◆ “10% Rule”: The sample size must not be bigger than 10% of the entire population.

(ii) Large Enough Sample Size:

◆ Sample size n should be large enough so that, $np \geq 10$ and $nq \geq 10$.

Example 7.6 Is CLT appropriate?

It is believed that nearsightedness affects about 8% of all children. 194 incoming children have their eyesight tested. Can the CLT be used in this situation?

Randomization: We have to assume there isn't some factor in the region that makes it more likely these kids have vision problems.

10% Rule: The population is “all children” — this is in the millions. 194 is less than 10% of the population.

Here, $np = 194 \times 0.08 = 15.52$, $nq = 194 \times 0.92 = 176.48$.

We have to make one assumption when using the CLT in this situation.

7.4.1 Central Limit Theorem (Sample Mean)

Theorem 1 Central Limit Theorem (Sample Mean)

If X_1, X_2, \dots, X_n are n random variables that are independent and identically distributed with mean μ and standard deviation σ . $\bar{x} = (X_1 + X_2 + \dots + X_n) / n$ is the sample mean, we can show

$$E(\bar{x}) = \mu \quad \text{and} \quad SD(\bar{x}) = \sigma / \sqrt{n}$$

then CLT states,

$$\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$$

as $n \rightarrow \infty$.

Implication of CLT

We have $\frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \rightarrow N(0,1)$, which means $\bar{x} \rightarrow N(\mu, \sigma^2 / n)$. So the sample mean can be approximated with a normal random variable with mean μ and standard deviation $\sigma \sqrt{n}$.

Example 7.7 Proportions of a Sample

Let's say we have a population with probability p of a certain characteristic (and $q = 1 - p$). We have a random sample of n from the population. What is the mean and standard deviation of the proportion of our sample that has the characteristic?

We can use the CLT if n is large enough. If X is the number of times the characteristic is found in our sample, $\tilde{p} = X/n$, our sample proportion, has mean p and standard deviation $\sqrt{(pq/n)}$.

7.4.2 Central Limit Theorem (Sample Sum)

Theorem 1 Central Limit Theorem (Sample Sum)

If X_1, X_2, \dots, X_n are n random variables that are independent and identically distributed with mean μ and standard deviation σ . $S_n = X_1 + X_2 + \dots + X_n$ is the sample sum, we can show

$$E(S_n) = n\mu \quad \text{and} \quad SD(S_n) = \sigma\sqrt{n}$$

then CLT states,

$$\frac{S - n\mu}{\sigma\sqrt{n}} \rightarrow N(0,1)$$

as $n \rightarrow \infty$.

Applications of CLT

Example 7.8 (Example 7.7 continued)

It is believed that nearsightedness affects about 8% of all children. 194 incoming children have their eyesight tested. Can the CLT be used in this situation?

\bar{x} should be approximately normally distributed have a mean of $0.08 \times 194 = 15.52$ and SD of $\sqrt{0.08 \times 0.92 \times 194} = 3.7787$, see Figure 7.14.

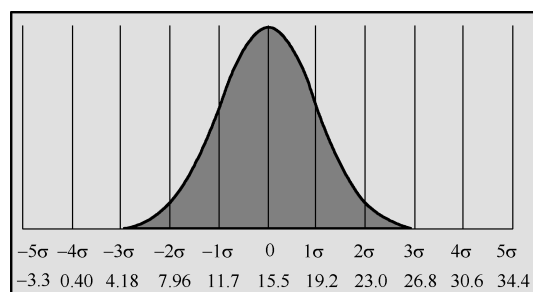


Figure 7.14 Distributions of Sample Means

Example 7.9 (CLT for Proportions)

How is the proportion of nearsighted children distributed? Divide by $n=194$: mean is 0.08, SD is 0.0195, see Figure 7.15.

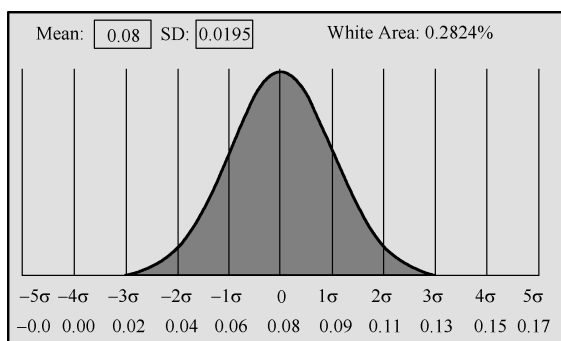


Figure 7.15 Distributions of Sample Means

Example 7.10 We flip 400 coins, assigning 1 for head and 0 for tail each time. What's the probability that the average outcome will be greater than or equal to 0.57?

If X is the outcome for a single coin and X_{400} is the average outcome of 400 coins then for X we have

$$E(X) = 1/2(1) + 1/2(0) = 1/2$$

$$\sigma(X) = \sqrt{\text{Var}(X)} = \sqrt{1/2(1-1/2)^2 + 1/2(0-1/2)^2} = 1/2$$

The CLT says that the corresponding approximating normal distribution has

$$\mu = 1/2$$

$$\sigma = (1/2) / \sqrt{400} = 0.025$$

In friendly terms

400-Coin Flip Average Outcome Distribution

≈ Normal Distribution with $\mu = 1/2$ and $\sigma = 0.025$

So our desired value is $P(0.57 \leq X_{400} < \infty)$ but by the CLT we can approximate this with the normal distribution. We find the corresponding z-scores

$$x = \infty \Rightarrow z = \infty$$

$$x = 0.57 \Rightarrow z = (0.57 - 0.5) / 0.025 = 2.8$$

and so

$$P(0.57 \leq X_{400} < \infty) \approx P(2.8 \leq Z < \infty) \approx 1 - 0.9974 = 0.0026 = 0.26\%$$

This means that the probability of getting an average outcome of over 0.57 from flipping 400 coins is approximately 0.26%.

Note: Limit theorems, in particular, the central limit theorems (CLT), surely are among the most important theorems in probability theory and statistics. They play an essential role in various applied sciences as well, including statistical mechanics.

The mathematics which prove the Central Limit Theorem are beyond the scope of this book, so we will not discuss them here.

Central Limit Theorem

In probability theory, the central limit theorem (CLT) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution. To illustrate what this means, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (commonly known as a “bell curve”). A simple example of this is that if one flips a coin many times, the probability of getting a given number of heads should follow a normal curve, with mean equal to half the total number of flips.

The central limit theorem has a number of variants. In its common form, the random variables must be identically distributed. In variants, convergence of the mean to the normal distribution also occurs for non-identical distributions or for non-independent observations, given that they comply with certain conditions.

The central limit theorem has an interesting history. The first version of this theorem was postulated by the French-born mathematician Abraham de Moivre who, in a remarkable article published in 1733, used the normal distribution to approximate the distribution of the number of heads resulting from many tosses of a fair coin. This finding was far ahead of its time, and was nearly forgotten until the famous French mathematician Pierre-Simon Laplace rescued it from obscurity in his monumental work *Théorie Analytique des Probabilités*, which was published in 1812. Laplace expanded De Moivre's finding by approximating the binomial distribution with the normal distribution. But as with De Moivre, Laplace's finding received little attention in his own time. It was not until the nineteenth century was at an end that the importance of the central limit theorem was discerned, when, in 1901, Russian mathematician Aleksandr Lyapunov defined it in general terms and proved precisely how it worked mathematically. Nowadays, the central limit theorem is considered to be the unofficial sovereign of probability theory.

The actual term “central limit theorem” (in German: “zentraler Grenzwertsatz”) was first used by George Pólya in 1920 in the title of a paper. Pólya referred to the theorem as “central” due to its importance in probability theory. According to Le Cam, the French school of probability interprets the word central in the sense that “it describes the behaviour of the centre of the distribution as opposed to its tails”.

Historically A. de Moivre, P.S. de Laplace, S.D. Poisson and C.F. Gauss have first shown that Gaussian is the attractor of independent systems with a finite second variance. Chebyshev, Markov, Liapounov, Feller, Lindeberg, Levy have contributed essentially to the development of the central limit theorem.

Problems

7.1 Manufacturers use random samples to test whether or not their product is meeting specifications. These samples could be people, manufactured parts, or even samples during the manufacturing of potato chips.

- Do you think that all random samples taken from the same population will lead to the same result?
- What characteristic (or property) of random samples could be observed during the sampling process?

7.2 Consider the set of odd single-digit integers $\{1, 3, 5, 7, 9\}$.

a. Make a list of all samples of size 2 that can be drawn from this set of integers. (Sample with replacement; that is, the first number is drawn, observed, and then replaced [returned to the sample set] before the next drawing.)

b. Construct the sampling distribution of sample means for samples of size 2 selected from this set.

c. Construct the sampling distributions of sample ranges for samples of size 2.

7.3 Consider the set of even single-digit integers $\{0, 2, 4, 6, 8\}$.

a. Make a list of all the possible samples of size 3 that can be drawn from this set of integers. (Sample with replacement; that is, the first number is drawn, observed, and then replaced [returned to the sample set] before the next drawing.)

b. Construct the sampling distribution of the sample medians for samples of size 3.

c. Construct the sampling distribution of the sample means for samples of size 3.

7.4 According to The World Factbook (《世界概括》, 又译作《世界各国纪实年鉴》; ISSN 1553-8133), 2004, the total fertility rate (estimated mean number of children born per woman) for Madagascar(马达加斯加) is 5.7. Suppose that the standard deviation of the total fertility rate is 2.6. The mean number of children for a sample of 200 randomly selected women is one value of many that form the SDSM.

- What is the mean value for this sampling distribution?
- What is the standard deviation of this sampling distribution?
- Describe the shape of this sampling distribution.

7.5 The USDA Economics and Statistics System at Cornell University maintains a Poultry Yearbook in which they list monthly, quarterly, and annual facts about the poultry industry. The 2004 yearbook lists the annual consumption of turkey meat as 17.71 pounds per person. Suppose the standard deviation for the consumption of turkey per person is 6.3 pounds. The mean weight of turkey consumed for a sample of 150 randomly selected people is one value of many that form the SDSM.

- What is the mean value for this sampling distribution?
- What is the standard deviation of this sampling distribution?
- Describe the shape of this sampling distribution.

7.6 Consider a normal population with $\mu = 43$ and $\sigma = 5.2$. Calculate the z -score for an 2 of 46.5 from a sample of size 16.

7.7 Consider a population with $\mu = 43$ and $\sigma = 5.2$.

- Calculate the z -score for an \bar{x} of 46.5 from a sample of size 35.
- Could this z -score be used in calculating probabilities using Table 3 in Appendix B? Why or why not?

7.8 What is the probability that the sample of kindergarten children (Example 7.5) has a mean height of less than 39.75 inches?

7.9 The local bakery bakes more than a thousand 1-pound loaves of bread daily, and the weights of these loaves varies. The mean weight is 1 lb, and 1 oz., or 482 grams. Assume the standard deviation of the weights is 18 grams and a sample of 40 loaves is to be randomly selected.

- This sample of 40 has a mean value of \bar{x} , which belongs to a sampling distribution. Find the shape of this sampling distribution.
- Find the mean of this sampling distribution.
- Find the standard error of this sampling distribution.
- What is the probability that this sample mean will be between 475 and 495 grams?
- What is the probability that the sample mean will have a value less than 478 grams?
- What is the probability that the sample mean will be within 5 grams of the mean?

7.10 Consider the approximately normal population of heights of male college students with mean $\mu = 69$ inches and standard deviation $\sigma = 4$ inches. A random sample of 16 heights is obtained.

- Describe the distribution of x , height of male college students.
- Find the proportion of male college students whose height is greater than 70 inches.
- Describe the distribution of \bar{x} , the mean of samples of size 16.
- Find the mean and standard error of the distribution.
- Find $P(\bar{x} > 70)$.
- Find $P(\bar{x} < 67)$.

7.11 Suppose the random variable X corresponds to rolling a die with winnings and losings given as follows: Rolling a 1 wins \$2, rolling a 2 or 3 wins \$7, rolling a 4 wins nothing and rolling a 5 or 6 loses \$6. Suppose we roll 2000 dice and average our winnings from each roll. What's the probability that we'll win under \$0.75? (Hit, using CLT)

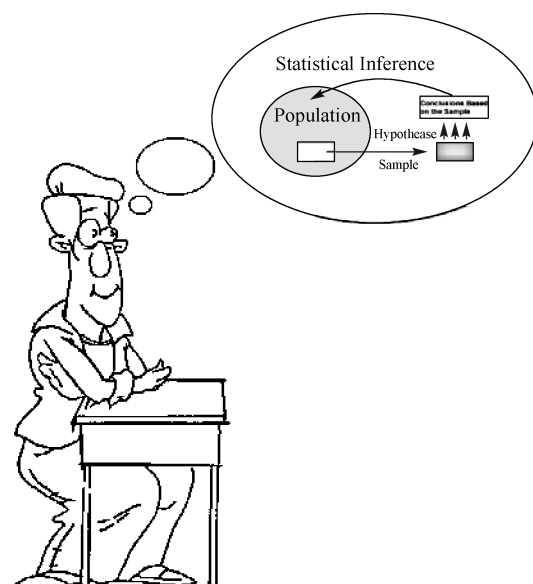
7.12 The lifespan (寿命) of a bacterium (细菌) is randomly distributed with mean 5 days and standard deviation 2 days. Suppose you have 160 bacteria in your lab. What's the probability that the average lifespan will be over 4.8 days?

Part II Inferential Statistics

A statistical model is a probability distribution constructed to enable inferences to be drawn or decisions made from data.

—A. C. Davison

(Department of Mathematics , Swiss Federal Institute of Technology)



Unit 8

Introduction to Statistical Inferences



8.1 Point Estimation and Interval Estimation



8.2 Estimation of Mean μ (σ Known)



8.3 Introduction to Hypothesis Testing



8.4 Formulating the Statistical Null and Alternative Hypotheses



8.5 Hypothesis Test of Mean μ (σ Known): A Probability–Value Approach



8.6 Hypothesis Test of Mean μ (σ Known): A Classical Approach



Problems

The objective of inferential statistics is to use the information contained in the sample data to increase our knowledge of the sampled population. We will learn about making two types of inferences : (i) estimating the value of a population parameter and (ii) testing a hypothesis. The sampling distribution of sample means (SDSM) is the key to making these inferences.

In this unit, we deal with questions about the population mean using two methods that assume that the value of the population standard deviation is a known quantity. This assumption is seldom realized in real-life problems. But it will make our first look at the techniques of inference much simpler.

8.1 Point Estimation and Interval Estimation

8.1.1 Point Estimate

Estimations occur in two forms: a point estimate and interval estimate.

A **point estimate for a parameter** is a single number designed to estimate a quantitative parameter of a population, usually the value of the corresponding sample statistic.

Example 8.1 Shearing strength

Shearing strength is the force required to break a material in a “cutting” action, see Figure 8.1.

To illustrate this, let’s look at a company that manufactures rivets for use in building aircraft. One characteristic of extreme importance is the “shearing strength” of each rivet. The company’s engineers must monitor production to be certain that the shearing strength of the rivets meets the required specs. To accomplish this, they take a sample and determine the mean shearing strength of the sample. Based on this sample information, the company can estimate the mean shearing strength for all the rivets it is manufacturing.

A random sample of 36 rivets is selected, and each rivet is tested for shearing strength. The resulting sample mean is $\bar{x}=924.23$ lb. Based on this sample, we say, “We believe the mean shearing strength of all such rivets is 924.23 lb.” That is, the sample mean, \bar{x} , is the point estimate (single number value) for the mean μ of the sampled population. For our rivet example, 924.23 is the point estimate for μ , the mean shearing strength of all rivets.

Note: Throughout Unit 8 we will treat the standard deviation σ , as a known, or given quantity and concentrate on learning the procedures for making statistical inferences about the population mean μ . Therefore, to continue the explanation of statistical inferences, we will assume $\sigma=18$ for the specific rivets described in our example.

Definition 1

■ **Point estimate for a parameter:** A single number designed to estimate a quantitative parameter of a population, usually the value of the corresponding sample statistic.

Sample means vary in value and form a sampling distribution in which not all samples result in \bar{x} values equal to the population mean. Therefore, we should not expect this sample of 36 rivets to produce a point estimate (sample mean) that is exactly equal to the mean μ of the sampled population. We should, however, expect the point estimate to be fairly close in value of the

population mean. The sampling distribution of sample means (SDSM) and the central limit theorem (CLT) provide the information needed to describe how close the point estimate, \bar{x} , is expected to be to the population mean, μ .

Recall that approximately 95% of a normal distribution is within two standard deviations of the mean and that the central limit theorem describes the sampling distribution of sample means as being nearly normal when samples are large enough. Samples of size 36 from populations of variables like rivet strength are generally considered large enough. Therefore, we should anticipate that approximately 95% of all random samples selected from a population with unknown mean μ and standard deviation $\sigma=18$ will have means \bar{x} between

$$\begin{aligned} &\mu - 2(\sigma_x) \text{ and } \mu + 2(\sigma_x) \\ &\mu - 2\left(\frac{\sigma}{\sqrt{n}}\right) \text{ and } \mu + 2\left(\frac{\sigma}{\sqrt{n}}\right) \\ &\mu - 2\left(\frac{18}{\sqrt{36}}\right) \text{ and } \mu + 2\left(\frac{18}{\sqrt{36}}\right) \end{aligned}$$

$$\mu - 6 \text{ and } \mu + 6$$

This suggests that 95% of all random samples of size 36 selected from the population of rivets should have a mean \bar{x} between $\mu - 6$ and $\mu + 6$. Figure 8.1 shows the middle 95% of the distribution, the bounds of the interval covering the 95%, and the mean μ .

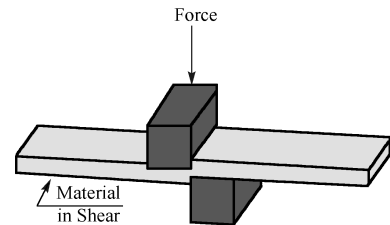


Figure 8.1 The shear strength of the material has to be

8.1.2 Interval Estimate

The second form of estimate is an **interval estimate**, which is an interval bounded by two values and used to estimate the value of a population parameter. The values that bound this interval are statistics calculated from the sample that is being used as the basis for the estimation. Interval estimates involve a certain level of confidence, $1-\alpha$, which is the proportion of all interval estimates that include the parameter being estimated.

Definition 2

■ **Interval estimate:** An interval bounded by two values and used to estimate the value of a population Parameter. The values that bound this interval are statistics calculated from the sample that is being used as the basis for the estimation.

Combining an interval estimate with a specified level of confidence gives us a confidence interval. We can pull all of the information from our rivet example together in the form of a confidence interval. To construct the confidence interval, we will use the point estimate \bar{x} as the central value of an interval in much the same way as we used the mean μ as the central value to find the interval that captures the middle 95% of the \bar{x} distribution in Figure 8.2.

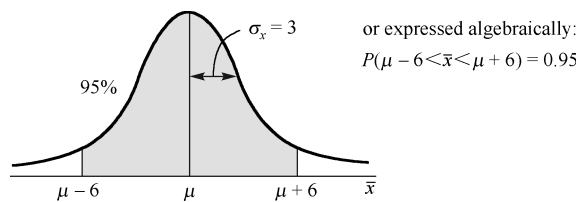


Figure 8.2 Sampling Distribution of \bar{x} 's Unknown μ

Definition 3

■ **Level of confidence $1-\alpha$:** The proportion of all interval estimates that include the parameter being estimated.

Definition 4

■ **Confidence interval:** An interval estimate with a specified level of confidence.

Example 8.2 (Example 8.1 continued)

For our rivet example, we can find the bounds to an interval centered at \bar{x} :

$$\begin{array}{ccc} \bar{x} - 2(\sigma_{\bar{x}}) & \text{to} & \bar{x} + 2(\sigma_{\bar{x}}) \\ 924.23 - 6 & \text{to} & 924.23 + 6 \end{array}$$

The resulting interval is 918.23 to 930.23

The level of confidence assigned to this interval is approximately 95%, or 0.95. The bounds of the interval are 2 multiples ($z = 2.0$) of the standard error from the sample mean, and by looking at Statistical Table 1 in Appendix, we can more accurately determine the level of confidence as 0.9544. Putting all of this information together, we express the estimate as a confidence interval: 918.23 to 930.23 is the 95.44% confidence interval for the mean shear strength of the rivets. Or in an abbreviated form: 918.23 to 930.23, the 95.44% confidence interval for μ .

New Words and Expressions

cut [kʌt] *vt. & vi.* 将 (某物) 切开 (或分割)

rivet ['rivɪt] *n.* 铆钉 *vt.* 铆接; 把.....固定住; 加深 (爱情, 友谊等)

specs [speks] *n.* 眼镜; 投机, 说明, 规格 (spec 的名词复数); 规范

anticipate [æn'tɪsɪpeɪt] *vt.* 预感; 预见; 预料

pull [pʊl] *vt.* 赢得; 吸引异性; (耍手腕) 得逞; 拉; 拖 *n.* 拖; 爬; 影响力

Technical Terms

point estimate 点估计

interval estimate 区间估计

level of confidence 置信水平

confidence interval 置信区间

8.2 Estimation of Mean μ (σ Known)

8.2.1 The Principle of Constructing a Confidence Interval

Information from the sampling distribution of sample means and the central limit theorem is used in estimating the value of an unknown **population mean**.

In section 8.1 we surveyed the basic ideas of estimation: point estimate, interval estimate, level of confidence, and confidence interval. These basic ideas are interrelated and used throughout statistics when an inference calls for an estimate. In this section we formalize the interval estimation process as it applies to estimating the population mean μ based on a random sample under the restriction that the population standard deviation σ is a known value.

The sampling distribution of sample means and the central limit theorem provide us with the information we need to ensure that the necessary assumptions are satisfied.

Definition 5

■ **The assumption for estimating mean μ using a known σ :** The sampling distribution of \bar{x} has a normal distribution.

The information needed to ensure that this assumption (or condition) is satisfied is contained in the sampling distribution of sample means (SDSM) and in the central limit theorem. Recall from Unit 7 that the sampling distribution of sample means \bar{x} is distributed about a mean equal to μ with a standard error equal to σ/\sqrt{n} ; and (1) if the randomly sampled population is normally distributed, then \bar{x} is normally distributed for all sample sizes, or (2) if the randomly sampled population is not normally distributed, then \bar{x} is approximately normally distributed for sufficiently large sample sizes.

Therefore, we can satisfy the required assumption by either (i) knowing that the sampled population is normally distributed or (ii) using a random sample that contains a sufficiently large number of data. The first possibility is obvious. We either know enough about the population to know that it is normally distributed or we don't. The second way to satisfy the assumption is by applying the CLT. Inspection of various graphic displays of the sample data should yield an indication of the type of distribution the population possesses. The CLT can be applied to smaller samples (say, $n = 15$ or larger) when the data provide a strong indication of a unimodal distribution that is approximately symmetric. If there is evidence of some skewness in the data, then the sample size needs to be much larger (perhaps $n \geq 50$). If the data provide evidence of an extremely skewed or J-shaped distribution, the CLT will still apply if the sample is large enough. In extreme cases, "large enough" may be unrealistically or impractically large. There is no hard-and-fast rule defining "large enough"; the sample size that is "large enough" varies greatly according to the distribution of the population.

Note: The word *assumptions* is somewhat of a misnomer. It does not mean that we "assume"

something to be the situation and continue, but that we must be sure the conditions expressed by the assumptions do exist before we apply a particular statistical method.

The $1 - \alpha$ confidence interval for the estimation of mean μ is found using the formula

$$\bar{x} - z(\alpha/2) \left(\frac{\sigma}{\sqrt{n}} \right) \quad \text{to} \quad \bar{x} + z(\alpha/2) \left(\frac{\sigma}{\sqrt{n}} \right) \quad (8.1)$$

Sometimes, $z(\alpha/2)$ is referred to as $z_{(\alpha/2)}$.

Here are the parts of the confidence interval formula:

(1) \bar{x} is the point estimate and the center point of the confidence interval.

(2) $z(\alpha/2)$ or $z_{(\alpha/2)}$ is the **confidence coefficient**.

It is the number of multiples of the standard error needed to formulate an interval estimate of the correct width to have a level of confidence of $1 - \alpha$. Figure 8.3

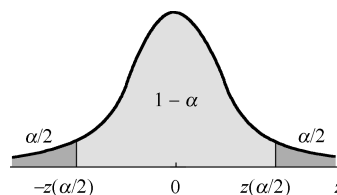


Figure 8.3 Confidence Coefficient $z(\alpha/2)$

shows the relationship among the level of confidence $1 - \alpha$ (the middle portion of the distribution), $(\alpha/2)$ (the “area to the right” used with the critical-value notation), and the confidence coefficient $z(\alpha/2)$ (whose value is found using Table 2(II) of Appendix Tables).

(3) $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean, or the standard deviation of the sampling distribution of sample means.

(4) $z(\alpha/2) \left(\frac{\sigma}{\sqrt{n}} \right)$ is one-half the width of the confidence interval (the product of the confidence coefficient and the standard error) and is called the maximum error of estimate, E .

(5) $\bar{x} - z(\alpha/2) \left(\frac{\sigma}{\sqrt{n}} \right)$ is called the lower confidence limit (LCL), and $\bar{x} + z(\alpha/2) \left(\frac{\sigma}{\sqrt{n}} \right)$ is called the upper confidence limit (UCL) for the confidence interval.

The estimation procedure is organized into a five-step process that will take into account all of the above information and produce both the point estimate and the confidence interval.

Basically, the confidence interval is “point estimate \pm maximum error”.

◇Summary: Find the Confidence Interval: A Five-Step Procedure ◇

Step 1 The Set-Up: Describe the population parameter of interest.

Step 2 The Confidence Interval Criteria:

- Check the assumptions.
- Identify the probability distribution and the formula to be used.
- State the level of confidence, $1 - \alpha$.

Step 3 The Sample Evidence: Collect the sample information.

Step 4 The Confidence Interval:

- Determine the confidence coefficient.

- b. Find the maximum error of estimate.
- c. Find the lower and upper confidence limits.

Step 5 The Results:

State the confidence interval.

8.2.2 Applications

Example 8.3

Let's apply the confidence interval procedure to finding the mean for a one-way commute distance. The student body at many community colleges is considered a "commute population". The student activities office wishes to obtain an answer to the question: How far (one way) does the average community college student commute to college each day?

Typically the "average student's commute distance" is meant to be the "mean distance" commuted by all students who commute. A random sample of 100 commuting students was identified, and the one-way distance each commuted was obtained. The resulting sample mean distance was 10.22 miles. To estimate the mean one-way distance commuted by all commuting students, we'll use: (i) a point estimate and (ii) a 95% confidence interval. (Use $\sigma = 6$ miles.) Our point estimate (a) for the mean one-way distance is 10.22 miles (the sample mean). Next we use the five-step procedure to find the 95% confidence interval (b).

Step 1 The Set-Up: Describe the population parameter of interest.

The mean μ of the one-way distances commuted by all commuting community college students is the parameter of interest.

Step 2 The Confidence Interval Criteria:

a. Check the assumptions.

Here σ is known. The variable "distance commuted" most likely has a skewed distribution because the vast majority of the students will commute between 0 and 25 miles, with fewer commuting more than 25 miles. A sample size of 100 should be large enough for the CLT to satisfy the assumption; the \bar{x} sampling distribution is approximately normal.

b. Identify the probability distribution and the formula to be used.

The standard normal distribution, z , will be used to determine the confidence coefficient, and formula (8.1) with $\sigma = 6$.

c. State the level of confidence, $1 - \alpha$.

The question asks for 95% confidence, or $1 - \alpha = 0.95$.

Step 3 The Sample Evidence: Collect the sample information.

The sample information is given in the statement of the problem: $n = 100$, $\bar{x} = 10.22$.

Step 4 The Confidence Interval:

a. Determine the confidence coefficient.

The confidence coefficient is found using Table 8.1:

Table 8.1 A Portion of Table 2(II)

Level of coefficient: $1 - \alpha = 0.95$	A Portion of Table 2(II)			Confidence confidence: $z(\alpha / 2) = 1.96$
	α	...	0.05	
	$z(\alpha / 2)$...	1.96	
	$1 - \alpha$...	0.95	

b. Find the maximum error of estimate.

Use the maximum error part of formula (8.1):

$$E = z(\alpha / 2) \left(\frac{\sigma}{\sqrt{n}} \right) = 1.96 \times 0.6 = 1.176$$

c. Find the lower and upper confidence limits.

Using the point estimate, \bar{x} , from Step 3 and the maximum error, E , from Step 4b, we find the confidence interval limits:

$$\begin{aligned} \bar{x} - z(\alpha / 2) \left(\frac{\sigma}{\sqrt{n}} \right) \quad \text{to} \quad \bar{x} + z(\alpha / 2) \left(\frac{\sigma}{\sqrt{n}} \right) \\ 10.22 - 1.176 \quad \text{to} \quad 10.22 + 1.176 \\ 9.044 \quad \text{to} \quad 11.396 \\ 9.04 \quad \text{to} \quad 11.40 \end{aligned}$$

Step 5 The Results: State the confidence interval.

9.04 to 11.40, the 95% confidence interval for μ . That is, with 95% confidence we can say, “The mean one-way distance is between 9.04 and 11.40 miles”, see Figure 8.4.

Let’s take another look at the concept “level of confidence”. It was defined to be the probability that the sample to be selected will produce interval bounds that contain the parameter.

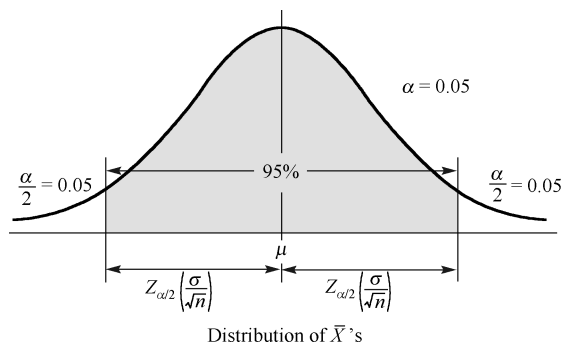


Figure 8.4 95% confidence interval of the mean

8.2.3 Sample Size and Confidence Interval

The confidence interval has two basic characteristics that determine its quality: its level of confidence and its width. It is preferred that the interval has a high level of confidence and be precise (narrow) at the same time. The higher the level of confidence, the more likely the interval is to contain the parameter, and the narrower the interval, the more precise the estimation.

However, these two properties seem to work against each other, since it would seem that a narrower interval would tend to have a lower probability and a wider interval would be less precise. The maximum error part of the confidence interval formula specifies the relationship involved, see Figure 8.5.

$$\bar{x} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Figure 8.5 Confidence Interval and Confidence level

Maximum Error of Estimate

$$E = z(\alpha / 2) \left(\frac{\sigma}{\sqrt{n}} \right) \quad (8.2)$$

This formula has four components: (1) the maximum error E , half of the width of the confidence interval; (2) the confidence coefficient, $z(\alpha/2)$, which is determined by the level of confidence; (3) the sample size, n ; and (4) the standard deviation, σ . The standard deviation σ is not a concern in this discussion because it is a constant (the standard deviation of a population does not change in value). That leaves three factors. Inspection of formula (8.2) indicates the following: Increasing the level of confidence will make the confidence coefficient larger and will thereby require either the maximum error to increase or the sample size to increase; decreasing the maximum error will require the level of confidence to decrease or the sample size to increase; and decreasing the sample size will force the maximum error to increase or the level of confidence to decrease. We have a “three-way tug of war”, as pictured in Figure 8.6.

An increase or decrease to any one of the three factors has an effect on one or both of the other two factors. The statistician’s job is to “balance” the level of confidence, the sample size, and the maximum error so that an acceptable interval results.

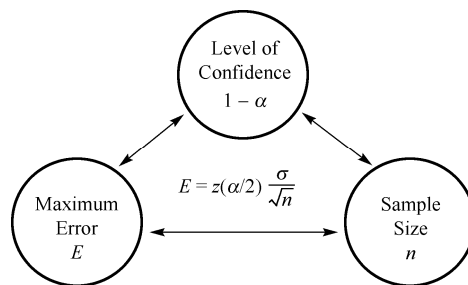


Figure 8.6 The “Three-Way Tug-of-War” between $1 - \alpha$, n , and E

Sample Size and Confidence Interval

To get a better idea of how statisticians balance confidence, sample size, and error, let’s look at the problem of determining the size sample needed to estimate the mean weight of all second-grade boys and to be accurate within 1 lb with We’ll assume a normal distribution and that the standard deviation of the boys’ weights is 3 lb.

The desired level of confidence determines the confidence coefficient: The confidence coefficient is found using Statistical Table 2(II): $z(\alpha / 2) = z(0.025) = 1.96$.

We know our desired maximum error is $E = 1.0$ (remember, 1 lb). Now we are ready to use the maximum error formula:

$$E = z(\alpha / 2) \left(\frac{\sigma}{\sqrt{N}} \right); \quad 1.0 = 1.96 \left(\frac{3}{\sqrt{n}} \right)$$

$$\text{Solve for } n: \quad 1.0 = \frac{5.88}{\sqrt{n}}$$

$$\sqrt{n} = 5.88, \quad n = (5.88)^2 = 34.57 = 35$$

Therefore, $n = 35$ is the sample size needed if you want a 95% confidence interval with a maximum error no greater than 1 lb.

Note: When we solve for the sample size n , it is customary to round up to the next larger integer, no matter what fraction (or decimal) results.

Calculating Sample Size with Unknown Value of Sigma (σ)

Using the maximum error formula (8.2) can be made a little easier by rewriting the formula in a form that expresses n in terms of the other values.

So, sample size formula is following:

$$n = \left(\frac{z(\alpha / 2) \cdot \sigma}{E} \right)^2 \quad (8.3)$$

If the maximum error is expressed as a multiple of the standard deviation σ , then the actual value of σ is not needed in order to calculate the sample size. If we wanted to find the sample size needed to estimate the population mean to within 1/5 of a standard deviation with 99% confidence, we would first need to determine the confidence coefficient (using Statistical Table 2(II)): $1 - \alpha = 0.99$, $z(\alpha / 2) = 2.58$. The desired maximum error is $E = \frac{\sigma}{5}$.

Now we are ready to use the sample size formula (8.3):

$$\begin{aligned} n &= \left(\frac{z(\alpha / 2) \cdot \sigma}{E} \right)^2 \\ n &= \left(\frac{(2.58) \cdot \sigma}{\sigma / 5} \right)^2 = (2.58 \times 5)^2 \\ &= 12.90^2 = 166.41 \approx 167 \end{aligned}$$

New Words and Expressions

hard-and-fast ['hɑ:dn'fɑ:st] *adj.* 不可违逆的, 必须遵守的

balance ['bæləns] *vt.* 结平(账目); 使(在某物上)保持平衡; 使(各部分)协调

n. 平衡; 平衡力; (酿酒配料的)均衡

tug of war [tʌg ɒv wɔ:] *n.* 拔河, 两派间的激烈竞争。three-way tug of war 三路之战

lb (LB) 是英国和美国的重量单位“磅”的简写, 一磅=0.45359 千克, 即 453.59 克

Technical Terms

lower confidence limit (LCL) 置信下限

upper confidence limit (UCL) 置信上限

8.3 Introduction to Hypothesis Testing

To test a claim we must formulate a null hypothesis and alternative hypothesis.

We all make decisions every day of our lives. Some of these decisions are of major importance; others are seemingly insignificant. All decisions follow the same basic pattern. We weigh the alternatives; then, based on our beliefs and preferences and whatever evidence is available, we arrive at a decision and take the appropriate action. The statistical hypothesis test follows much the same process, except that it involves statistical information.

In this section we develop many of the concepts and attitudes of the hypothesis test while looking at several decision-making situations without using any statistics. To test claim we must formulate a null hypothesis and an alternative hypothesis.

8.3.1 Null Hypothesis and Alternative Hypothesis

Example 8.4

A friend is having a party (Super Bowl party, home-from-college party—you know the situation, any excuse will do), and you have been invited. You must make a decision: attend or not attend. That's simple; well maybe, except that you want to go only if you can be convinced the party is going to be more fun than your friend's typical party; furthermore, you definitely do not want to go if the party is going to be just another dud. You have taken the position that "the party will be a dud" and you will not go unless you become convinced otherwise. Your friend assures you, "Guaranteed, the party will be a great time!" Do you go or not?

The decision-making process starts by identifying *something of concern* and then formulating two hypotheses about it. A **hypothesis** is a statement that something is true. Your friend's statement, "The party will be a great time," is a hypothesis. Your position, "The party will be a dud," is also a hypothesis. The process by which a decision is made between two opposing hypotheses is called a **statistical hypothesis test**. The two opposing hypotheses are formulated so that each hypothesis is the negation of the other. (That way one of them is always true, and the other one is always false.) Then one hypothesis is tested in hopes that it can be shown to be a very improbable occurrence, thereby implying that the other hypothesis is likely the truth.

Definition 6

■ **Hypothesis:** A statement that something is true.

Definition 7

- **Statistical hypothesis test:** A process by which a decision is made between two opposing hypotheses.

Definition 8

- **Null hypothesis, H_o :** The hypothesis we will test.

Definition 9

- **Alternative hypothesis, H_a :** A statement about the same population parameter that is used in null hypothesis.

The two hypotheses involved in making a decision are known as the *null hypothesis* and the *alternative hypothesis*. The null hypothesis is the hypothesis we will test. Generally this is a statement that a population parameter has a specific value. The null hypothesis is so named because it is the “starting point” for the investigation. (The phrase “there is no difference” is often used in its interpretation.) The **alternative hypothesis** is a statement about the same population parameter that is used in the null hypothesis. Generally this is a statement that specifies that the population parameter has a value different, in some way, from the value given in the null hypothesis. The rejection of the null hypothesis will imply the likely truth of this alternative hypothesis.

With regard to your friend’s party, the two opposing viewpoints or hypotheses are: “The party will be a great time” and “The party will be a dud”. Which statement becomes the null hypothesis, and which becomes the alternative hypothesis?

Determining the statement of the null hypothesis and the statement of the alternative hypothesis is a very important step. The *basic idea* of the hypothesis test is for the evidence to have a chance to “disprove” the null hypothesis. The null hypothesis is the statement that the evidence might disprove. Your concern (belief or desired outcome), as the person doing the testing, is expressed in the alternative hypothesis. As the person making the decision, you believe that the evidence will demonstrate the feasibility of your “theory” by demonstrating the *unlikeliness* of the truth of the null hypothesis. The alternative hypothesis is sometimes referred to as the research hypothesis, since it represents what the researcher hopes will be found to be “true”. (If so, he or she will get a paper out of the research.)

We use the notation H_o for the null hypothesis to contrast it with H_a for the alternative hypothesis. Other texts may use H_o (subscript zero) in place of H_o and H_1 in place of H_a .

Since the “evidence” (who’s going to the party, what is going to be served, and so on) can only demonstrate the unlikeliness of the party being a dud, your initial position, “The party will be a dud,” becomes the null hypothesis. Your friend’s claim, “The party will be a great time, then becomes the alternative hypothesis.”

H_o : “Party will be a dud”

vs.

H_a : “Party will be a great time”

Here are some examples to give you some ideas of how to do this in different situations.

Example 8.5

You suspect that a brand-name detergent outperforms the store's brand of detergent, and you wish to test the two detergents because you would prefer to buy the cheaper store brand. State the null and alternative hypotheses.

Solution

Your suspicion, "the brand-name detergent outperforms the store brand." is the reason for the test and therefore becomes the alternative hypothesis.

H_a : "There is no difference in detergent performance."

H_o : "The brand-name detergent performs better than the store brand."

However, as a consumer, you are hoping not to reject the null hypothesis for budgetary reasons.

Example 8.6

You are testing a new design for airbags used in automobiles, and you are concerned that they might not open properly. State the null and alternative hypotheses.

Solution



The two opposing possibilities are "Bags open properly" and "Bags do not open properly". Testing can only produce evidence that discredits the hypothesis "Bags open properly". Therefore, the null hypothesis is "Bags open properly" and the alternative hypothesis is "Bags do not open properly."

The alternative hypothesis can be the statement the experimenter wants to show to be true.

8.3.2 Four Possible Outcomes in a Hypothesis Test

Before returning to our example about the party, we need to look at the four possible outcomes that could result from the null hypothesis being either true or false and the decision being either to "reject H_o " or to "fail to reject H_o ". Table 8.2 shows these four possible outcomes.

Table 8.2 Four Possible Outcomes in a Hypothesis Test

Decision	Null Hypothesis	
	True	False
Fail to reject H_o	Type A correct decision 	Type II error 'False negative'
Reject H_o	Type I error 'False positive'	Type B correct decision 

A **type A correct decision** occurs when the null hypothesis is true, and we decide in its favor. A **type B correct decision** occurs when the null hypothesis is false, and the decision is in opposition to the null hypothesis. A **type I error** is committed when a true null hypothesis is rejected—that is, when the null hypothesis is true but we decide against it. A **type II error** is committed when we decide in favor of a null hypothesis that is actually false.

When a decision is made, it would be nice to always make the correct decision. This, however, is not possible in statistics because we make our decisions on the basis of sample information. The best

we can hope for is to control the probability with which an error occurs. The probability assigned to the type I error is α (called “alpha”; α is the first letter of the Greek alphabet). The probability of the type II error is β (called “beta”; β is the second letter of the Greek alphabet), see Table 8.3.

Table 8.3 Probability with Which Decisions Occur

Error in Decision	Type	Probability	Correct Decision	Type	Probability
Rejection of a true H_o	I	α	Failure to reject a true H_o	A	$1 - \alpha$
Failure to reject a false H_o	II	β	Rejection of a false H_o	B	$1 - \beta$

Notes:

(i) The truth of the situation is not known before the decision is made, the conclusion reached, and the resulting actions take place. The truth of H_o may never be known.

(ii) The type II error often results in what represents a “lost opportunity”; lost in this situation is the chance to use a product that yields better results.

To control these errors, we will assign a small probability to each of them. The most frequently used probability values for α and β are 0.01 and 0.05. The probability assigned to each error depends on its seriousness. The more serious the error, the less willing we are to have it occur, and therefore a smaller probability will be assigned, α and β are probabilities of errors, each under separate conditions, and they cannot be combined. Therefore, we cannot determine a single probability for making an incorrect decision. Likewise, the two correct decisions are distinctly separate and each has its own probability; $1 - \alpha$ is the probability of a correct decision when the null hypothesis is true, and $1 - \beta$ is the probability of a correct decision when the null hypothesis is false. $1 - \beta$ is called the power of the statistical test, since it is the measure of the ability of a hypothesis test to reject a false null hypothesis, a very important characteristic.

Example 8.5 (Continued) Describing Outcomes

How would we describe the four possible outcomes and the resulting actions that would occur for the hypothesis test about detergent described example 8.5.

First, recall that we need a hypothesis and a null hypothesis, see Table 8.4.

H_o : “There is no difference in detergent performance”

H_a : “The brand-name detergent performs better than the store brand”

Table 8.4 Null Hypothesis about Detergent Described Example 8.5

	Null Hypothesis Is True	Null Hypothesis Is False
Fail to Reject H_o	<p><i>Type A Correct Decision</i> Truth of situation: There is no difference between the deterclents. Conclusion: It was determined that there was no difference. Action: The consumer bought the cheaper detergent, saving money and getting the same results</p>	<p><i>Type II Error</i> Truth of situation: The brand-name detergent is better. Conclusion: It was determined that there was no difference. Action: The consumer bought the cheaper detergent, saving money and getting inferior results</p>
Reject H_o	<p><i>Type I Error</i> Truth of situation: There is no difference between the detergents. Conclusion: It was determined that the brand-name detergent was better. Action: The consumer bought the brand-name detergent, spending extra money to attain no better results</p>	<p><i>Type B Correct Decision</i> Truth of situation: The brand-name detergent is better. Conclusion: It was determined that the brand-name detergent was better. Action: The consumer bought the brand-name detergent, spending more and getting better results</p>

Note: Regardless of the outcome of a hypothesis test, you are never certain that a correct decision has been reached.

Let's look back at the two possible errors in decision that could occur in our example about the laundry detergent. Most people would become upset if they found out they were spending extra money for a detergent that performed no better than the cheaper brand. Likewise, most people would become upset if they found out they could have been buying a better detergent. Evaluating the relative seriousness of these errors requires knowing whether this is your personal laundry or a professional laundry business, how much extra the brand-name detergent costs, and so on.

Type I Error, Type II Error and Sample Size

There is an interrelationship among the probability of the type I error (α), the probability of the type II error (β), and the sample size (n). This is very much like the interrelationship among level of confidence, maximum error, and sample size discussed in section 8.2.

Figure 8.7 shows the “three-way tug of war” among α , β , and n . If any one of the three is increased or decreased, it has an effect on one or both of the others. The statistician's job is to “balance” the three values of α , β , and n to achieve an acceptable testing situation. If α is reduced, then either β must increase or n must be increased; if β is decreased, then either α increases or n must be increased; if n is decreased, then either α increases or β increases. The choices for α , β , and n are definitely

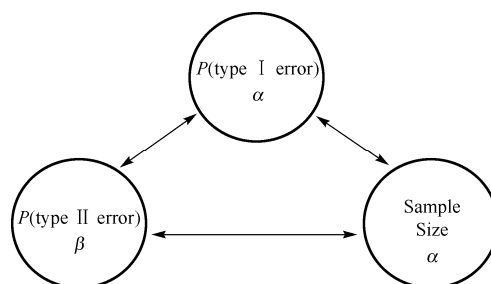


Figure 8.7 The “Three-Way Tug-of-War” between α , β , and n

not arbitrary. At this time in our study of statistics, α will be given in the statement of the problem, as will the sample size n . Further discussion on the role of β , $P(\text{type II error})$, is left for another time.

Sample size, n , is self-explanatory, so let's look at the role of α , or the level of significance. The level of significance α is the probability of committing the type I error.

Establishing the level of significance can be thought of as a “managerial decision”. Typically, someone in charge determines the level of probability with which he or she is willing to risk a type I error.

At this point in the hypothesis test procedure, the evidence is collected and summarized and the value of a test statistic is calculated. A test statistic is a random variable whose value is calculated from the sample data and is used in making the decision “fail to reject H_0 ” or “reject H_0 ”. The value of the calculated test statistic is used in conjunction with a decision rule to determine either “reject H_0 ” or “fail to reject H_0 ”. This decision rule must be established prior to collecting the data; it specifies how you will reach the decision.

Back to your friend's party: You have to weigh the history of your friend's parties, the time and place, others going, and so on, against your own criteria and then make your decision. As a result of the decision about the null hypothesis (“The party will be a dud”), you will take the appropriate action; you will either go to or not go to the party.

To complete a hypothesis test, you will need to write a conclusion that carefully describes the meaning of the decision relative to the intent of the hypothesis test.

When writing the decision and the conclusion, remember that (1) the decision is about H_o and (2) the conclusion is a statement about whether or not the contention of H_o was upheld. This is consistent with the “attitude” of the whole hypothesis test procedure. The null hypothesis is the statement that is “on trial”, and therefore the decision must be about it. The contention of the alternative hypothesis is the thought that brought about the need for a decision. Therefore, the question that led to the alternative hypothesis must be answered when the conclusion is written.

We must always remember that when the decision is made, nothing has been proved. Both decisions can lead to errors: “Fail to reject H_o ” could be a type II error (the lack of sufficient evidence has led to great parties being missed more than once), and “reject H_o ” could be a type I error (more than one person has decided to go to a party that was a dud).

The Conclusion

(i) If the decision is “reject H_o ”, then the conclusion should be worded something like, “There is sufficient evidence at the a level of significance to show that... (the meaning of the alternative hypothesis).”

(ii) If the decision is “fail to reject H_o ”, then the conclusion should be worded something like, “There is not sufficient evidence at the a level of significance to show that... (the meaning of the alternative hypothesis).”

New Words and Expressions

weigh [weɪ] *vt.* 称.....的重量；权衡，考虑；用手掂估

vi. 具有重要性；重量为；成为.....的重荷

dud [dʌd] *adj.* 无用的；不完善的 *n.* 无用的人或物；衣服，个人物品

assure [ə'ʃʊə(r)] *vt.* 向.....保证；使.....确信

disprove [ˌdɪs'pru:v] *vt.* 反驳；证明.....是虚假的

unlikeliness [ʌn'laɪklɪnəs] *n.* 不大可能，不一样

brand-name 著名品牌的

laundry ['lɔ:ndri] *n.* 洗衣店，洗衣房；洗好的衣服；待洗的衣服

detergent [dɪ'tɜ:dʒənt] *n.* 洗涤剂；去垢剂

upset [ʌp'set] *vt.* 打乱，搅乱；推翻，弄翻；使心烦意乱；使翻倒

n. 翻倒，颠覆；心烦意乱；混乱

interrelationship [ˌɪntərɪ'leɪʃnʃɪp] *n.* 相互关系 [联系，影响]，干扰

risk [rɪsk] *vt.* 冒.....的危险；使.....冒风险（或面临危险）

self-explanatory [self ɪk'splænətɪ] *adj.* 不解自明的，明显的；自解释

Notes

同义词辨析 :ensure, insure, assure, guarantee, pledge, promise 这些动词都有“ 保证 ”之意。

ensure : 侧重使人相信某个行为或力量产生的结果。

insure : insure 常与 ensure 换用, 但前者多指经济方面的保证、保险。

assure : 侧重指消除某人思想上的怀疑或担心, 从而有达到目的的保证感, 但不如 ensure 普通。

guarantee : 指对事物的品质或人的行为及履行义务、义务等承担责任的保证。

pledge : 正式用词, 指通过郑重许诺、协议或立誓等保证承担某一义务或遵守某一原则。

promise : 侧重表达自己的主观意向, 设法用语言使人感到稳当可靠。

8.4 Formulating the Statistical Null and Alternative Hypotheses

Let's look at two examples that demonstrate formulating the statistical null and alternative hypotheses involving the population mean μ .

The trichotomy law from algebra states that two numerical values must be related in exactly one of three possible relationships: $<$, $=$, or $>$. All three of these possibilities must be accounted for in the two opposing hypotheses in order for the two hypotheses to be negations of each other. The three possible combinations of signs and hypotheses are shown in Table 8.5. Recall that the null hypothesis assigns a specific value to the parameter in question, and therefore “equals” will always be part of the null hypothesis.

Table 8.5 The Three Possible Statements of Null and Alternative Hypotheses

Null Hypothesis	Alternative Hypothesis
1. greater than or equal to (\geq)	less than ($<$)
2. less than or equal to (\leq)	greater than ($>$)
3. equal to ($=$)	not equal to (\neq)

8.4.1 Writing Null and Alternative Hypothesis in One-Tailed Situation

Example 8.7

Suppose the EPA was suing the city of Rochester for noncompliance with carbon monoxide standards. Specifically, the EPA would want to show that the mean level of carbon monoxide in downtown Rochester's air is dangerously high, higher than 4.9 parts per million. State the null and alternative hypotheses.

Solution

To state the two hypotheses, we first need to identify the population parameter in question: the “mean level of carbon monoxide in Rochester.” The parameter μ is being compared to the value 4.9 parts per million, the specific value of interest. The EPA is questioning the value of μ and wishes to show that it is higher than 4.9 (that is, $\mu > 4.9$). The three possible relationships—(1) $\mu <$

4.9, (2) $\mu = 4.9$, and (3) $\mu > 4.9$ —must be arranged to form two opposing statements: One states the EPA’s position, “The mean level is higher than 4.9 ($\mu > 4.9$),” and the other states the negation, “The mean level is not higher than 4.9 ($\mu \leq 4.9$).” One of these two statements will become the null hypothesis H_o , and the other will become the alternative hypothesis H_a .

Note: Recall that there are two rules for forming the hypotheses: (1) The null hypothesis states that the parameter in question has a specified value (“ H_o must contain the equal sign”), and (2) the EPA’s contention becomes the alternative hypothesis (“higher than”). Both rules indicate:

$$H_o: \mu = 4.9 (\leq) \quad \text{and} \quad H_a: \mu > 4.9$$

8.4.2 Writing Null and Alternative Hypothesis in Two-Tailed Situation

Example 8.8

Job satisfaction is very important to worker productivity. A standard job-satisfaction questionnaire was administered by union officers to a sample of assembly line workers in a large plant in hopes of showing that the assembly workers’ mean score on this questionnaire would be different from the established mean of 68. State the null and alternative hypotheses.

Solution

Either the mean job satisfaction score is different from 68 ($\mu \neq 68$) or the mean is equal to 68 ($\mu = 68$). Therefore,

$$H_o: \mu = 68 \quad \text{and} \quad H_a: \mu \neq 68$$

Table 8.6 lists some additional common phrases used in claims and indicates their negations and the hypothesis in which each phrase will be used. Again, notice that “equals” is always in the null hypothesis. Also notice that the negation of “less than” is “greater than or equal to”. Think of negation as “all the others” from the set of three signs.

Table 8.6 Common Phrases and Their Negations

$H_o:(\geq)$	$H_o:(<)$	$H_o:(\leq)$	$H_o:(>)$	$H_o:(=)$	$H_o:(\neq)$
at least	less than	at most	more than	is	is not
no less than	less than	no more than	more than	not different from	different from
not less than	less than	not greater than	greater than	the same as	not the same as

After the null and alternative hypotheses are established, we will work under the assumption that the null hypothesis is a true statement until there is sufficient evidence to reject it. This situation might be compared to a courtroom trial, where the accused is assumed to be innocent (H_o : Defendant is innocent vs. H_a : Defendant is not innocent) until sufficient evidence has been presented to show that innocence is totally unbelievable (“beyond reasonable doubt”). At the conclusion of the hypothesis test, we will make one of two possible decisions. We will decide in opposition to the null hypothesis and say that we “reject H_o ” (this corresponds to “conviction” of the accused in a trial), or we will decide in agreement with the null hypothesis and say that we “fail to reject H_o ” (this corresponds to “fail to convict” or an “acquittal” of the accused in a trial).

New Words and Expressions

- trichotomy [traɪ'kɒtəmi] *n.* 三分法
noncompliance [ˌnɒnkəm'plaɪəns] *n.* 不服从, 不顺从; 不履行
monoxide [mɒ'nɒksaɪd] *n.* 一氧化物
downtown [ˌdaʊn'taʊn] *n.* 市中心区; (市中) 商业区
dangerously ['deɪndʒərəsli] *adv.* 危险地, 可能引起危险地
defendant [dɪ'fendənt] *n.* [法]被告人 *adj.* [法]被告的; 辩护的
conviction [kən'vɪkʃn] *n.* 定罪; 说服; 确信

Notes

trichotomy law 三分律

8.5 Hypothesis Test of Mean μ (σ Known): A Probability-Value Approach

In section 8.3 and 8.4, we surveyed the concepts and much of the reasoning behind a hypothesis test while looking at nonstatistical illustrations. In this section, we are going to formalize the hypothesis test procedure as it applies to statements concerning the mean μ of a population under the restriction that σ , the population standard deviation, is a known value.

The assumption for hypothesis tests about mean μ using a known σ : The sampling distribution of \bar{x} has a normal distribution.

The information we need to ensure that this assumption is satisfied is contained in the sampling distribution of sample means and in the central limit theorem (see Unit 7):

The sampling distribution of sample means \bar{x} is distributed about a mean equal to μ with a standard error equal to σ/\sqrt{n} ; and (1) if the randomly sampled population is normally distributed, then \bar{x} is normally distributed for all sample sizes, or (2) if the randomly sampled population is not normally distributed, then \bar{x} is approximately normally distributed for sufficiently large sample sizes.

The hypothesis test is a well-organized, step-by-step procedure used to make a decision. Two different formats are commonly used for hypothesis testing. The probability-value approach, or simply p -value approach, is the hypothesis test process that has gained popularity in recent years. The p -value approach makes full use of computer's capability in the work of the decision-making process. This approach is organized as a five-step procedure outlined in the box below.

◇The Probability-Value Hypothesis Test: A Five-Step Procedure ◇

Step 1 The Set-Up:

- Describe the population parameter of interest.
- State the null hypothesis (H_0) and the alternative hypothesis (H_a).

Step 2 The Hypothesis Test Criteria:

- a. Check the assumptions.
- b. Identify the probability distribution and the test statistic to be used.
- c. Determine the level of significance, σ .

Step 3 The Sample Evidence:

- a. Collect the sample information.
- b. Calculate the value of the test statistic.

Step 4 The Probability Distribution:

- a. Calculate the p -value for the test statistic.
- b. Determine whether or not the p -value is smaller than α .

Step 5 The Results:

- a. State the decision about H_0 .
- b. State the conclusion about H_0 .

8.5.1 One-Tailed Hypothesis Test Using the p -Value Approach

Example 8.9

To get a sense of how this procedure works, let's consider a commercial aircraft manufacturer that buys rivets to use in assembling airliners. Each rivet supplier that wants to sell rivets to the aircraft manufacturer must demonstrate that its rivets meet the required specifications. One of the specs is: "The mean shearing strength of all such rivets, μ , is at least 925 lb." Each time the aircraft manufacturer buys rivets, it is concerned that the mean strength might be less than the 925 lb specification.

Each individual rivet has a shearing strength, which is determined by measuring the force required to shear ("break") the rivet. Clearly, not all the rivets can be tested. Therefore, a sample of rivets will be tested, and a decision about the mean strength of all the untested rivets will be based on the mean from those sampled and tested.

Step 1 The Set-Up:

a. Describe the population parameter of interest. The population parameter of interest is the mean μ , the mean shearing strength of (or mean force required to shear) the rivets being considered for purchase.

b. State the null hypothesis (H_0) and the alternative hypothesis (H_a).

The null hypothesis and the alternative hypothesis are formulated by inspecting the problem or statement to be investigated and first formulating two opposing statements about the mean μ . For our example, these two opposing statements are: (A) "The mean shearing strength is less than 925" ($\mu < 925$, the aircraft manufacturer's concern), and (B) "The mean shearing strength is at least 925" ($\mu = 925$, the rivet supplier's claim and the aircraft manufacturer's spec).

The parameter of interest, the population mean μ , is related to the value 925. Statement (A) becomes the alternative hypothesis:

$$H_a: \mu < 925 \text{ (the mean is less than 925)}$$

This statement represents the aircraft manufacturer's concern and says, "The rivets do not meet the required specs." Statement (B) becomes the null hypothesis:

$$H_o: \mu = 925 (\geq) \text{ (the mean is at least 925)}$$

This hypothesis represents the negation of the aircraft manufacturer's concern and says, "The rivets do meet the required specs."

Note: We will write the null hypothesis with just the equal sign, thereby stating the exact value assigned. When "equal" is paired with "less than" or paired with "greater than", the combined symbol is written beside the null hypothesis as a reminder that all three signs have been accounted for in these two opposing statements.

Step 2 The Hypothesis Criteria:

a. Check the assumptions.

σ is known. Variables like shearing strength typically have a mounded distribution; therefore, a sample of size 50 should be large enough for the CLT to apply and ensure that the sampling distribution of sample means will be normally distributed.

b. Identify the probability distribution and the test statistic to be used.

The standard normal probability distribution is used because \bar{x} is expected to have a normal distribution.

For a hypothesis test of μ , we want to compare the value of the sample mean to the value of the population mean as stated in the null hypothesis. This comparison is accomplished using the test statistic in formula (8.4):

Test Statistic for Mean

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \quad (8.4)$$

The resulting calculated value is identified as z ("z star") because it is expected to have a standard normal distribution when the null hypothesis is true and the assumptions have been satisfied. The ("star") is to remind us that this is the calculated value of the test statistic.

The test statistic to be used is $z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ with $\sigma = 18$.

c. Determine the level of significance, α

Setting α was described as a managerial decision in section 8.3. To see what is involved in determining α , the probability of the type I error, for our rivet example, we start by identifying the four possible outcomes, their meaning, and the action related to each.

The type I error occurs when a true null hypothesis is rejected. This would occur when the manufacturer tested rivets that in truth did meet the specs, and rejected them. Undoubtedly this would lead to the rivets not being purchased even though they did meet the specs. In order for the manager to set a level of significance, related information is needed—namely, how soon is the new supply of rivets needed? If they are needed tomorrow and this is the only vendor with an available supply, waiting a week to find acceptable rivets could be very expensive; therefore, rejecting good rivets could be considered a serious error. On the other hand, if the rivets are not needed until next month, then this error may not be very serious. Only the manager will know all the ramifications, and therefore the manager's input is important here.

After much consideration, the manager assigns the level of significance: $\alpha = 0.05$.

Step 3 The Sample Evidence:

a. Collect the sample information.

We are ready for the data. The sample must be a random sample drawn from the population whose mean μ is being questioned. A random sample of 50 rivets is selected, each rivet is tested, and the sample mean shearing strength is calculated: $\bar{x} = 921.18$ and $n = 50$.

b. Calculate the value of the test statistic.

The sample evidence (\bar{x} and n found in Step 3a) is next converted into the calculated value of the test statistic, z , using formula (8.4). (μ is 925 from H_0 ; $\sigma = 18$ is the known quantity, as shown to the left.) We have

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}; \quad z = \frac{921.18 - 925.0}{18 / \sqrt{50}} \\ = \frac{-3.82}{2.5456} = -1.50$$

Step 4 The Probability Distribution:

a. Calculate the p -value for the test statistic. The probability value, or p -value, is the probability that the test statistic could be the value it is or a more extreme value (in the direction of the alternative hypothesis) when the null hypothesis is true. The p -value is represented by the area under the curve of the probability distribution for the test statistic that is more extreme than the calculated value of the test statistic. There are three separate cases, and the direction (or sign) of the alternative hypothesis is the key. Table 8.7 outlines the procedure for all three cases.

Table 8.7 Finding p -Values

Case 1 H_a contains “>” “Right tail”	p -value is the area to right of z $p\text{-value} = P(z > z^*)$	<p>p-Value in Right Tails</p>
Case 2 H_a contains “<” “Left tail”	p -value is the area to left of z the area of the left tail is the same as the area in the right tail bounded by the positive z ; therefore, $p\text{-value} = P(z < z^*) = P(z > z^*)$	<p>p-value in Two Tails</p>
Case 3 H_a contains “≠” “Two-tailed”	p -value is the total area of both tails $p\text{-value} = P(z < - z^*) + P(z > z^*)$ z^* may be in either tail, and since both areas are equal, find the probability of one tail and double it. Thus, $p\text{-value} = 2 \times P(z > z^*)$	<p>p-value in Two Tails</p>

To apply this to Step 4 of our rivet example, draw a sketch of the standard normal distribution and locate z (found in Step 3b, see Figure 8.8) on it. To identify the area that represents the p -value, look at the sign in the alternative hypothesis. For this test, the alternative hypothesis indicates that we are interested in that part of the sampling distribution that is “less than” z . Therefore, the p -value is the area that lies to the left of z . Shade this area. You can see that for this example, we are dealing with a one-tailed, left tail hypothesis.

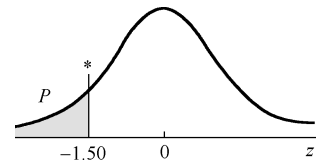


Figure 8.8 There are three ways to find

Method 1: Use Statistical Table 1 in Appendix to determine the tabled area related to $z = 1.50$; then calculate the p -value by subtracting from 0.5000:

$$\begin{aligned} p\text{-value} &= P(z < z^*) = P(z < -1.50) = P(z > 1.50) \\ &= 0.5000 - 0.4332 = \mathbf{0.0668} \end{aligned}$$

Method 2: Use Statistical Table 3 in Appendix and the symmetry property: Table 3 is set up to allow you to read the p -value directly from the table. Since $P(z < -1.50) = P(z > 1.50)$, simply locate $z = 1.50$ on Table 5 and read the p -value:

$$P(z < -1.50) = \mathbf{0.0668}$$

Method 3: Use the cumulative probability function on a computer or calculator to find the p -value:

$$P(z < -1.50) = \mathbf{0.0668}$$

b. Determine whether or not the p -value is smaller than α .

In our example, the p -value (0.0668) is not smaller than α (0.05).

Step 5 The Results:

a. State the decision about H_o . Is the p -value small enough to indicate that the sample evidence is highly unlikely in the event that the null hypothesis is true? In order to make the decision, we need to know the decision rule.

Decision about H_o : Fail to reject H_o .

b. State the conclusion about H_a .

◇ **Decision Rule** ◇

(i) If the p -value is less than or equal to the level of significance α , then the decision must be to reject H_o .

(ii) If the p -value is greater than the level of significance α , then the decision must be to fail to reject H_o .

Review the conclusion of the Section 8.3. In our case, there is not sufficient evidence at the 0.05 level of significance to show that the mean shearing strength of the rivets is less than 925. We “failed to convict” the null hypothesis. In other words, a sample mean as small as 921.18 is likely to occur (as defined by α) when the true population mean value is 925.0 and \bar{x} is normally distributed. The resulting action by the manager would be to buy the rivets.

Note: When the decision reached is “fail to reject H_o ” (or “accept H_o ” as many say improperly), it simply means “for the lack of better information, act as if the null hypothesis is true”.

8.5.2 Two-Tailed Hypothesis Test Using the p -Value Approach

Let’s now look at an illustration involving the two-tailed procedure. To do this, we’ll use the example of an employee selection test.

Example 8.10

Many large companies in a certain city have for years used the Kelly Employment Agency for testing prospective employees. The employment selection test used has historically resulted in scores normally distributed about a mean of 82 and a standard deviation of 8. The Brown Agency has developed a new test that is quicker and easier to administer and therefore less expensive. Brown claims that its test results are the same as those obtained on the Kelly test. Many of the companies are considering a change from the Kelly Agency to the Brown Agency in order to cut costs. However, they are unwilling to make the change if the Brown test results have a different mean value. An independent testing firm tested 36 prospective employees with the Brown test. A sample mean of 79 resulted. Determine the p -value associated with this hypothesis test. (Assume $\sigma = 8$.)

Step 1 The Set-up:

a. Describe the population parameter of interest.

The population mean μ , the mean of all test scores using the Brown Agency test.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

The Brown Agency’s test results “will be different” (the concern) if the mean test score is not equal to 82. They “will be the same” if the mean is equal to 82. Therefore,

$$H_o: \mu = 82 \text{ (test results have the same mean)}$$

$$H_a: \mu \neq 82 \text{ (test results have a different mean)}$$

Step 2 The Hypothesis Test Criteria:

a. Check the assumptions.

σ is known. If the Brown test scores are distributed the same as the Kelly test scores, they will be normally distributed and the sampling distribution will be normal for all sample sizes.

b. Identify the probability distribution and the test statistic to be used.

The standard normal probability distribution and the test statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

will be used with $\sigma = 8$.

c. Determine the level of significance, α .

The level of significance is omitted because the question asks for the p -value and not a decision.

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 36$, $\bar{x} = 79$.

b. Calculate the value of the test statistic.

μ is 82 from H_0 ; $\sigma = 8$ is the known quantity.

We have

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}; \quad z = \frac{79 - 82}{8 / \sqrt{36}} = \frac{-3}{1.3333} = -2.25$$

Step 4 *The Probability Distribution:*

a. Calculate the p -value for the test statistic.

Since the alternative hypothesis indicates a two-tailed test, we must find the probability associated with both tails. The p -value is found by doubling the area of one tail (see Table 8.7).

Since $z = -2.25$, the value of $|z| = 2.25$.

The p -value $= 2 \times P(z > |z|)$

$$= 2 \times P(z > 2.25).$$

From Table 1: p -value $= 2 \times P(z > 2.25)$

$$= 2 \times (0.5000 - 0.4878)$$

$$= 2(0.0122)$$

$$= 0.0244.$$

1/2 p -value	table value
----------------	-------------

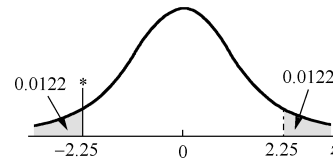


Figure 8.9 p -value for the test statistic

or

From Table 3: p -value $= 2 \times P(z > 2.25)$

$$= 2(0.0122)$$

$$= \mathbf{0.0244}.$$

or

Use the cumulative probability function on a computer or calculator.

b. Determine whether or not the p -value is smaller than α .

A comparison is not possible; no α value was given in the statement of the question.

Step 5 **The Results:**

The p -value for this hypothesis test is 0.0244. Each individual company now will decide whether to continue to use the Kelly Agency's services or change to the Brown Agency. Each will need to establish the level of significance that best fits its own situation and then make a decision using the decision rule described previously.

8.5.3 Evaluating the p -Value Approach

The fundamental idea of the p -value is to express the degree of belief in the null hypothesis:

- When the p -value is minuscule (something like 0.0003), the null hypothesis would be rejected by everybody because the sample results are very unlikely for a true H_0 .
- When the p -value is fairly small (like 0.012), the evidence against H_0 is quite strong and H_0 will be rejected by many.
- When the p -value begins to get larger (say, 0.02 to 0.08), there is too much probability that data like the sample involved could have occurred even if H_0 were true, and the rejection of H_0 is not an easy decision.

○ When the p -value gets large (like 0.15 or more), the data are not at all unlikely if the H_0 is true, and no one will reject H_0 .

The advantages of the p -value approach are as follows: (1) The results of the test procedure are expressed in terms of a continuous probability scale from 0.0 to 1.0, rather than simply on a “reject” or “fail to reject” basis. (2) A p -value can be reported and the user of the information can decide on the strength of the evidence as it applies to his or her own situation. (3) Computers can do all the calculations and report the p -value, thus eliminating the need for tables.

The disadvantage of the p -value approach is the tendency for people to put off determining the level of significance. This should not be allowed to happen, because it is then possible for someone to set the level of significance after the fact, leaving open the possibility that the “preferred” decision will result. This is probably important only when the reported p -value falls in the “hard choice” range (say, 0.02 to 0.08), as described previously.

New Words and Expressions

ramification [ˌræmɪfɪˈkeɪʃn] *n.* 衍生物, 结果; 分叉, 分支; 支流

improper [ɪmˈprɒpə(r)] *adj.* 不合适的, 非正常的; 不正确的; 不正派的

convict [kənˈvɪkt] *n.* 罪犯. *vt.* 宣判有罪; 证明……有罪; 定……的罪

8.6 Hypothesis Test of Mean μ (σ Known): A Classical Approach

The classical approach uses critical values in doing the work of the decision-making process.

In section 8.5, we explored the p -value approach to hypothesis testing. Now, we'll examine the classical approach, which has enjoyed popularity for many years. Like the p -value approach, the classical approach is a well-organized, step-by-step procedure used to make a decision. The classical hypothesis test is also organized as a five-step procedure.

With the classical procedure, we still assume that about mean μ using a known σ , the sampling distribution of \bar{x} has a normal distribution. (See section 8.5)

The Classical Hypothesis Test: A Five-Step Procedure

Step 1 The Set-Up:

- Describe the population parameter of interest.
- State the null hypothesis (H_0) and the alternative hypothesis (H_a).

Step 2 The Hypothesis Test Criteria:

- Check the assumptions.
- Identify the probability distribution and the test statistic to be used.
- Determine the level of significance, α .

Step 3 The Sample Evidence:

- a. Collect the sample information.
- b. Calculate the value of the test statistic.

Step 4 The Probability Distribution:

- a. Determine the critical region and critical value(s).
- b. Determine whether or not the calculated test statistic is in the critical region.

Step 5 The Results:

- a. State the decision about H_0 .
- b. State the conclusion about H_a .

8.6.1 One-Tailed Hypothesis Test Using the Classical Approach

Example 8.11 Writing Null and Alternative Hypothesis

A consumer advocate group would like to disprove a car manufacturer's claim that a specific model will average 24 miles per gallon of gasoline. Specifically, the group would like to show that the mean miles per gallon is considerably less than 24. State the null and alternative hypotheses.

Solution

To state the two hypotheses, we first need to identify the population parameter in question: the "mean mileage attained by this car model." The parameter μ is being compared to the value 24 miles per gallon, the specific value of interest. The advocates are questioning the value of μ and wish to show it to be less than 24 (that is, $\mu < 24$). There are three possible relationships: (1) $\mu < 24$, (2) $\mu = 24$, and (3) $\mu > 24$. These three cases must be arranged to form two opposing statements: One states what the advocates are trying to show, "The mean level is less than 24 ($\mu < 24$)," whereas the "negation" is "The mean level is not less than 24 ($\mu \geq 24$)." One of these two statements will become the null hypothesis H_0 , and the other will become the alternative hypothesis H_a .

Note: Recall that there are two rules for forming the hypotheses: (1) The null hypothesis states that the parameter in question has a specified value (" H_0 must contain the equal sign"), and (2) the consumer advocate group's contention becomes the alternative hypothesis ("less than"). Both rules indicate:

$$H_0 : \mu = 24(\geq) \quad \text{and} \quad H_a : \mu < 24$$

To keep things simple, let's return to our rivet example.

Example 8.12 (Example 8.1 continued)

Recall the setup: A commercial aircraft manufacturer buys rivets to use in assembling airliners. Each rivet supplier that wants to sell rivets to the aircraft manufacturer must demonstrate that its rivets meet the required specifications. One of the specs is: "The mean shearing strength of all such rivets, μ , is at least 925 lb." Each time the aircraft manufacturer buys rivets, it is concerned that the mean strength might be less than the 925-lb specification. The same stipulations about shearing strength and sampling apply here as well.

Step 1 The Set-Up:

As with the p -value approach, our basic problem set up is the same (see section 8.5.1). Note also that the trichotomy law from algebra is in force for the classical hypothesis test as well. The three possible combinations of signs and hypotheses were shown in Table 8.5. Recall that the null hypothesis assigns a specific value to the parameter in question, and therefore “equals” will always be part of the null hypothesis.

Again, the parameter of interest, the population mean μ , is related to the value 925. Statement (A) becomes the alternative hypothesis:

$$H_o : \mu < 925 \text{ (the mean is less than 925)}$$

This statement represents the aircraft manufacturer’s concern and says, “The rivets do not meet the required specs.” Statement (B) becomes the null hypothesis:

$$H_o : \mu = 925 (\geq) \text{ (the mean is at least 925)}$$

This hypothesis represents the negation of the air-craft manufacturer's concern and says, “The rivets do meet the required specs.”

Note: We use the same notation for writing the null hypothesis as we did with the p -value approach. (See section 8.5.1 for notes.)

Before continuing with our rivet example, let’s look at an example that demonstrates formulating the statistical null and alternative hypotheses involving population mean. Our example here is of a one-tailed situation. Writing for a two-tailed situation using the classical approach is comparable to writing it using the p -value approach (see section 8.5.1).

Whether you use the p -value or classical approach for the rivet example, Steps 2 and 3 are the same. To refresh your memory on how those were conducted, see section 8.5.1. When we get to Step 4, however, things are done a bit differently.

Step 4 The Probability Distribution:

a. Determine the critical region and critical value(s).

The standard normal variable z is our test statistic for this hypothesis test; therefore, we draw a sketch of the standard normal distribution, label the scale as z , and locate its mean value, 0. The **critical region** is the set of values for the test statistic that will cause us to reject the null hypothesis. The set of values that are not in the critical region is called the **noncritical region** (sometimes called the *acceptance region*).

Recall that we are working under the assumption that the null hypothesis is true. Thus, we are assuming that the mean shearing strength of all rivets in the sampled population is 925. If this is the case, then when we select a random sample of 50 rivets, we can expect this sample mean, \bar{x} , to be part of a normal distribution that is centered at 925 and to have a standard error of $\sigma/\sqrt{n} = 18/\sqrt{50}$, or approximately 2.55. Approximately 95% of the sample mean values will be greater than 920.8 [a value 1.65 standard errors below the mean: $925 - (1.65)(2.55) = 920.8$]. Thus, if H_o is true and $\mu = 925$, then we expect \bar{x} to be greater than 920.8 approximately 95% of the time and less than 920.8 only 5% of the time, see Figure 8.10.

If, however, the value of \bar{x} that we obtain from our sample is less than 920.8—say, 919.5—we will have to make a choice. It could be that either: (A) such a value (919.5) is a member of the sampling distribution with mean 925, although it has a very low probability of occurrence (less than 0.05), or (B) $\bar{x} = 919.5$ is a member of a sampling distribution whose mean is less than 925, which would make it a value that is more likely to occur, see Figure 8.11.

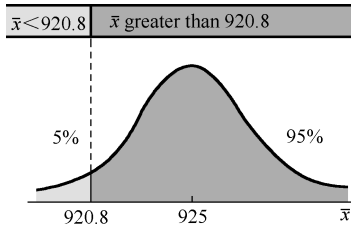


Figure 8.10 Probability be less than 920.8

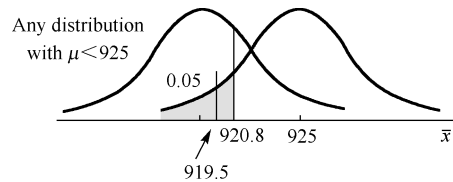


Figure 8.11 Sampling distribution that mean is less than 925

Definition 10

■ **Critical region:** The set of values for the test statistic that will cause us to reject the null hypothesis.

Definition 11

■ **Noncritical region (acceptance region):** The set of values that are not in the critical region.

Definition 12

■ **Critical value(s):** The “first” or “boundary” value(s) of the critical region(s).

In statistics, we “bet” on the “more probable to occur” and consider the second choice (B) to be the right one. Thus, the left-hand tail of the z -distribution becomes the critical region. And the level of significance α becomes the measure of its area.

We also need to identify the critical value, or first or boundary value, of the critical region. The critical value for our example is $-z(0.05)$ and has the value of -1.65 , as found in Statistical Table 2(I) in Appendix, see Figure 8.12.

b. Determine whether or not the calculated test statistic is in the critical region.

Graphically this determination is shown by locating the value for z on the sketch in Step 4a. The calculated value of z , $z = -1.50$, is not in the critical region (it is in the unshaded portion of the figure), see Figure 8.13.

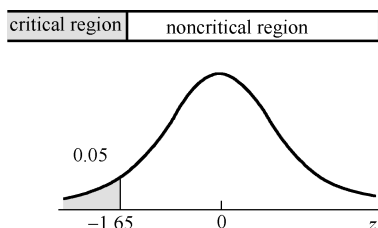


Figure 8.12 The critical value is $-z(0.05)$

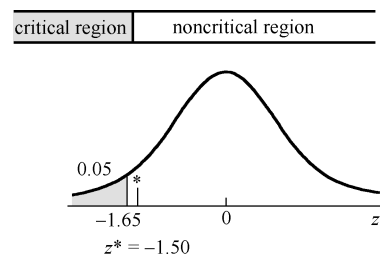


Figure 8.13 The area of $z = -1.50$

◇ **Decision Rule** ◇

- (i) If the test statistic falls within the critical region, then the decision must be to reject H_o . (The critical value is part of the critical region.)
- (ii) If the test statistic is not in the critical region, then the decision must be to fail to reject H_o .

Step 5 The Results:

a. State the decision about H_o .

In order to make the decision, we need to know the decision rule.

The decision is: Fail to reject H_o .

b. State the conclusion about H_a .

There is not sufficient evidence at the 0.05 level of significance to show that the rivets have a mean shearing strength less than 925. We “failed to convict” the null hypothesis. In other words, a sample mean as small as 921.18 is likely to occur (as defined by α) when the true population mean value is 925.0. Therefore, the resulting action would be to buy the rivets.

Let’s summarize briefly some of the details we have seen thus far:

- (i) Then null hypothesis specifies a particular value of a population parameter.
- (ii) The alternative hypothesis can take three forms. Each form dictates a specific location of the critical region(s), as shown in the following table.
- (iii) For many hypothesis tests, the sign in the alternative hypothesis “points” in the direction in which the critical region is located. Think of the not equal to sign (\neq) as being both less than ($<$) and greater than ($>$), thus pointing in both directions, see Table 8.8.

Table 8.8 Sign in the Alternative Hypothesis

	Sign in the Alternative Hypothesis		
	$<$	\neq	$>$
	One region	Two regions	One region
Critical Region	Left side	Half on each	Right side
	One-tailed test	Two-tailed test	One-tailed test

The value assigned to α is called the *significance level* of the hypothesis test. Alpha cannot be interpreted to be anything other than the risk (or probability) of rejecting the null hypothesis when it is actually true. We will seldom be able to determine whether the null hypothesis is true or false; we will decide only to “reject H_o ” or to “fail to reject H_o ”. The relative frequency with which we reject a true hypothesis is α , but we will never know the relative frequency with which we make an error in decision. The two ideas are quite different; that is, a type I error and an error in decision are two different things altogether. Remember that there are two types of errors: type I and type II.

8.6.2 Two-Tailed Hypothesis Test Using the Classical Approach

Let’s look at one last hypothesis test involving the two-tailed procedure.

Example 8.13 (Example 8.1 continued)

For this example, we’ll look at the mean weight of female college students. It has been

claimed that the mean weight of women students at a college is 54.4 kg. Professor Mark does not believe the claim and sets out to show that the mean weight is not 54.4 kg. To test the claim, he collects a random sample of 100 weights from among the women students. A sample mean of 53.75 kg results. To determine if this is sufficient evidence for Professor Mark to reject the statement, let $\alpha = 0.05$ and $\sigma = 5.4$ kg. Once again, we'll apply our five-step procedure.

Step 1 The Set-up:

a. Describe the population parameter of interest.

The population parameter of interest is the mean μ , the mean weight of all women students at the college.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

The mean weight is equal to 54.4 kg, or the mean weight is not equal to 54.4 kg.

$$H_o : \mu = 54.4 \text{ (mean weight is 54.4)}$$

$$H_a : \mu \neq 54.4 \text{ (mean weight is not 54.4)}$$

(Remember: \neq is $<$ and $>$ together.)

Step 2 The Hypothesis Test Criteria:

a. Check the assumptions.

σ is known. The weights of an adult group of women are generally approximately normally distributed; therefore, a sample of $n = 100$ is large enough to allow the CLT to apply.

b. Identify the probability distribution and the test statistic to be used.

The standard normal probability distribution and the test statistic

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

will be used; $\sigma = 5.4$.

c. Determine the level of significance, α .

$\alpha = 0.05$ (given in the statement of the problem).

Step 3 The Sample Evidence:

a. Collect the sample information:

$$\bar{x} = 53.75 \text{ and } n = 100.$$

b. Calculate the value of the test statistic.

Use formula (8.4), information from $H_o: \mu = 54.4$, and $\sigma = 5.4$ (known):

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}; \quad z^* = \frac{53.75 - 54.4}{5.4 / \sqrt{100}} = \frac{-0.65}{0.54} = -1.204 = -1.20$$

Step 4 The Probability Distribution:

a. Determine the critical region and critical value(s).

The critical region is both the left tail and the right tail because both smaller and larger values of the sample mean suggest that the null hypothesis is wrong. The level of significance will be split

in half, with 0.025 being the measure of each tail. The critical values are found in Statistical Table 2(II) in Appendix: $\pm z(0.025) = \pm 1.96$, see Figure 8.14. (Statistical Table 2(II) instructions are in Section 8.2)

b. Determine whether or not the calculated test statistic is in the critical region.

The calculated value of z , $z^* = -1.20$, is not in the critical region (shown in *black* on the figure above).

Step 5 The Results:

a. State the decision about H_0 : Fail to reject H_0 .

b. State the conclusion about H_a :

There is not sufficient evidence at the 0.05 level of significance to show that the women students have a mean weight different from the 54.4 kg claimed. In other words, there is no statistical evidence to support Professor Mark's contentions.

In this unit we have restricted our discussion of inferences to the mean of a population for which the standard deviation is known. In Units 9 and 10 we will discuss inferences about the population mean and remove the restriction about the known value for standard deviation. We will also look at inferences about the parameters proportion, variance, and standard deviation.

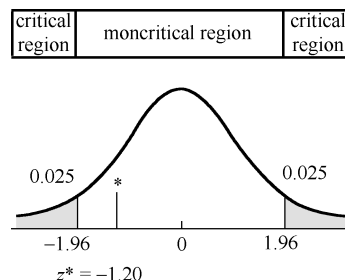


Figure 8.14 The area of $z^* = -1.50$

New Words and Expressions

advocate ['ædvəkət] *vt.* 提倡；拥护；为……辩护。 *n.* (辩护) 律师；提倡者；支持者
contention [kən'tenʃn] *n.* 竞争，争论；争夺；论点

Technical Terms

critical region 临界区域；拒绝域
noncritical region 非临界区域；非拒绝域
acceptance region 接受域

Problems

8.1 A random sample of the amount paid (in dollars) for taxi fare from downtown to the airport was obtained:

15 19 17 23 21 17 16 18 12 18 20 22 15 18 20

Use the data to find a point estimate for each of the following parameters.

a. Mean b. Variance c. Standard deviation

8.2 Find the level of confidence assigned to an interval estimate of the mean formed using the following intervals:

- a. $\bar{x} - 1.28 \cdot \sigma_{\bar{x}}$ to $\bar{x} + 1.28 \cdot \sigma_{\bar{x}}$
- b. $\bar{x} - 1.44 \cdot \sigma_{\bar{x}}$ to $\bar{x} + 1.44 \cdot \sigma_{\bar{x}}$
- c. $\bar{x} - 1.96 \cdot \sigma_{\bar{x}}$ to $\bar{x} + 1.96 \cdot \sigma_{\bar{x}}$
- d. $\bar{x} - 2.33 \cdot \sigma_{\bar{x}}$ to $\bar{x} + 2.33 \cdot \sigma_{\bar{x}}$

8.3 Determine the value of the confidence coefficient $z(\alpha/2)$ for each situation described:

- a. $1-\alpha = 0.90$
- b. $1-\alpha = 0.95$

8.4 Given the information, the sampled population is normally distributed, $n = 55$, $\bar{x} = 78.2$, and $\sigma = 12$:

- a. Find the 0.98 confidence interval for μ .
- b. Are the assumptions satisfied? Explain.

8.5 In your own words, describe the relationship between the following:

- a. Sample mean and point estimate
- b. Sample size, sample standard deviation, and standard error
- c. Standard error and maximum error

8.6 In your own words, describe the relationship between the point estimate, the level of confidence, the maximum error, and the confidence interval.

8.7 A sample of 60 night-school students' ages is obtained in order to estimate the mean age of night-school students. $\bar{x} = 25.3$ years. The population variance is 16.

- a. Give a point estimate for μ .
- b. Find the 95% confidence interval for μ .
- c. Find the 99% confidence interval for μ .

8.8 How large a sample should be taken if the population mean is to be estimated with 99% confidence to within \$75? The population has a standard deviation of \$900.

8.9 The new mini-laptop (迷你) computers can deliver as much computing power as machines several times their size, but they weigh in at less than 3 lb. How large a sample would be needed to estimate the population mean weight if the maximum error of estimate is to be 0.4 of 1 standard deviation with 95% confidence?

8.10 A supplier of highway construction materials claims he can supply an asphalt mixture that will make roads that are paved with his materials less slippery when wet. A general contractor who builds roads wishes to test the supplier's claim. The null hypothesis is "Roads paved with this asphalt mixture are no less slippery than roads paved with other asphalt." The alternative hypothesis is "Roads paved with this asphalt mixture are less slippery than roads paved with other asphalt."

- a. Describe the meaning of the two possible types of errors that can occur in the decision when this hypothesis test is completed.
- b. Describe how the null hypothesis, as stated previously, is a "starting point" for the decision to be made about the asphalt.

8.11 Describe the actions that would result in a type I error and a type II error if each of the following null hypotheses were tested. (Remember, the alternative hypothesis is the negation of the null hypothesis.)

- a. H_o : The majority of Americans favor laws against assault weapons.
- b. H_o : The choices on the fast-food menu are not low in salt.
- c. H_o : This building must not be demolished.
- d. H_o : There is no waste in government spending.

8.12 A normally distributed population is known to have a standard deviation of 5, but its mean is in question. It has been argued to be either $\mu = 80$ or $\mu = 90$, and the following hypothesis test has been devised to settle the argument. The null hypothesis, $H_o: \mu = 80$, will be tested using one randomly selected data value and comparing it with the critical value of 86. If the data value is greater than or equal to 86, the null hypothesis will be rejected.

- a. Find α , the probability of the type I error.
- b. Find β , the probability of the type II error.

8.13 State the null hypothesis H_o and the alternative hypothesis H_a that would be used for a hypothesis test related to each of the following statements:

- a. The mean age of the students enrolled in evening classes at a certain college is greater than 26 years.
- b. The mean weight of packages shipped on Air Express during the past month was less than 36.7 lb.
- c. The mean life of fluorescent light bulbs is at least 1600 hours.
- d. The mean strength of welds by a new process is different from 570 lb per unit area, the mean strength of welds by the old process.

8.14 Assume that z is the test statistic and calculate the value of z , for each of the following:

- a. $H_o: \mu = 51$, $\sigma = 4.5$, $n = 40$, $\bar{x} = 49.6$
- b. $H_o: \mu = 20$, $\sigma = 4.3$, $n = 75$, $\bar{x} = 21.2$
- c. $H_o: \mu = 138.5$, $\sigma = 3.7$, $n = 14$, $\bar{x} = 142.93$
- d. $H_o: \mu = 815$, $\sigma = 43.3$, $n = 60$, $\bar{x} = 799.6$

8.15 Calculate the p -value for each of the following:

- a. $H_o: \mu = 10$, $H_a: \mu > 10$, $z = 1.48$
- b. $H_o: \mu = 105$, $H_a: \mu < 105$, $z = -0.85$
- c. $H_o: \mu = 13.4$, $H_a: \mu \neq 13.4$, $z = 1.17$
- d. $H_o: \mu = 8.56$, $H_a: \mu < 8.56$, $z = -2.11$
- e. $H_o: \mu = 110$, $H_a: \mu \neq 110$, $z = -0.93$

8.16 State the null hypothesis, H_o , and the alternative hypothesis, H_a , that would be used for a hypothesis test for each of the following statements:

- a. The mean age of the youths who hang out at the mall is less than 16 years.
- b. The mean height of professional basketball players is greater than 6'6".
- c. The mean elevation drop for ski trails at eastern ski centers is at least 285 feet.
- d. The mean diameter of the rivets is no more than 0.375 inches.
- e. The mean cholesterol level of male college students is different from 200 mg/dL.

8.17 Determine the critical region and the critical values used to test the following null hypotheses:

- a. $H_o: \mu = 55 (\geq), H_o: \mu < 55, \alpha = 0.02$
- b. $H_o: \mu = -86 (\geq), H_o: \mu < -86, \alpha = 0.01$
- c. $H_o: \mu = 107, H_o: \mu \neq 107, \alpha = 0.05$
- d. $H_o: \mu = 17.4 (\leq), H_o: \mu > 17.4, \alpha = 0.10$

Statistics is a way of reasoning, along with a collection of tools and methods, designed to help us understand the world

— De Veaux et al.



Unit 9

Inferences Involving One Population



9.1 Inferences about the Mean μ (σ Unknown)



9.2 Inferences about the Binomial Probability of Success



9.3 Inferences about the Variance and Standard Deviation



Reading English Materials: The History of Statistics—Student's
t-statistic.



Problems

9.1 Inferences about the Mean μ (σ Unknown)

Inferences about the population mean μ are based on the sample mean \bar{x} and information obtained from the sampling distribution of sample means.

Recall that the sampling distribution of sample means has a mean μ and a standard error of σ/\sqrt{n} for all samples of size n , and it is normally distributed when the sampled population has a normal distribution or approximately normally distributed when the **sample size** is sufficiently large. This means that the test statistic $z^* = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ has a standard normal distribution. However, when σ is unknown, the standard error σ/\sqrt{n} is also unknown. Therefore, the sample standard deviation s will be used as the point estimate for σ . As a result, an estimated standard error of the mean, s/\sqrt{n} , will be used and our test statistic will become $\frac{\bar{x} - \mu}{s/\sqrt{n}}$.

When a known σ is being used to make an inference about the mean μ , a sample provides one value for use in the formulas; that one value is \bar{x} . When the sample standard deviation s is also used, the sample provides two values: the sample mean \bar{x} and the estimated standard error s/\sqrt{n} . As a result, the z -statistic will be replaced with a statistic that accounts for the use of an estimated standard error. This new statistic is known as the **Student's t -statistic**.

William Sealy Gosset and Student's t -statistic.

William Sealy Gosset (1876–1937) studied mathematics and chemistry at Oxford University and upon graduation took a position with Guinness Brewery in Dublin, where he found a mass of collected data related to the brewing process. In 1905, he met with *Karl Pearson* to discuss his statistical problems, and a year later, with Guinness' approval, he went to work at Pearson's Biometric Laboratory.

Upon returning to Guinness, he was put in charge of their Experimental Brewery. During these years he wrote several papers, which Guinness agreed to let him publish, provided he used a pseudonym and did not include company data; he thus used the pseudonym "A Student".

In 1908 W. S. Gosset, an Irish brewery employee, published a paper about this t -distribution under the pseudonym "Student". In deriving the t -distribution, Gosset assumed that the samples were taken from normal populations. Although this might seem to be restrictive, satisfactory results are obtained when large samples are selected from many nonnormal populations.

Figure 9.1 presents a diagrammatic organization for the inferences about the population mean as discussed in Unit 8 and in this first section of Unit 9. Two situations exist: σ is known, or σ is unknown. As stated before, σ is almost never a known quantity in real-world problems; therefore, the standard error will almost always be estimated by s/\sqrt{n} . The use of an estimated standard error of the mean requires the use of the t -distribution. Almost all real-world inferences about the population mean will be made with the Student's t -statistic.

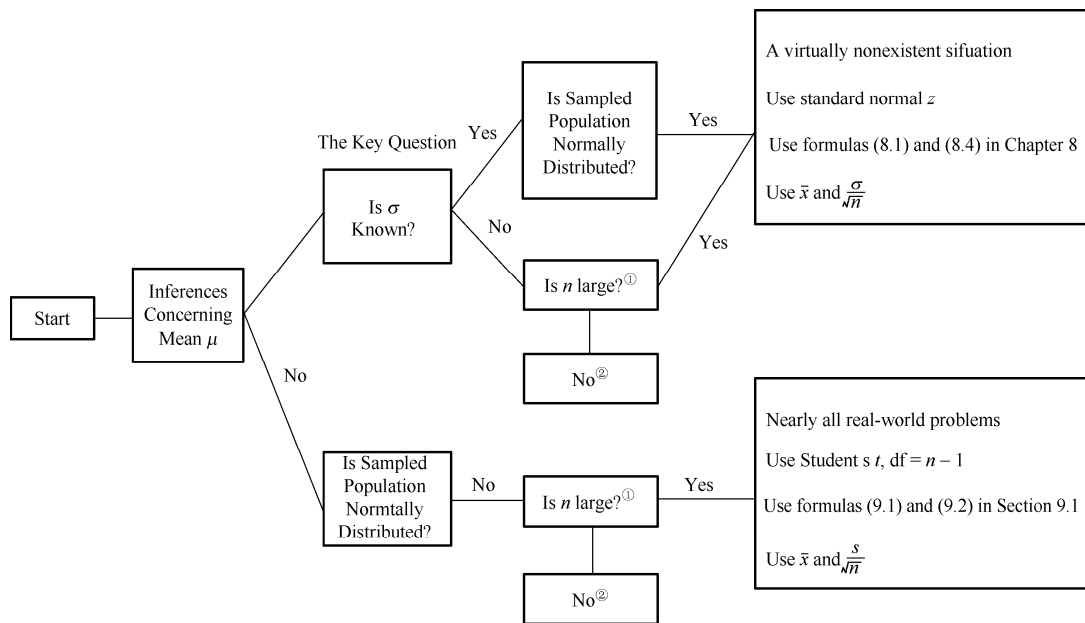


Figure 9.1 Do you use the z -statistic or the t -statistic

Is n large? Samples as small as $n = 15$ or 20 may be considered large enough for the central limit theorem to hold if the sample data are unimodal, nearly symmetrical, short-tailed, and without outliers. Samples that are not symmetrical require larger sample sizes, with 50 sufficing except for extremely skewed samples. See the discussion in Unit 8 (section 8.2).

Requires the use of a nonparametric technique; see advanced statistics or nonparametric statistics book.

The t -distribution has the following properties, see also Figure 9.2:

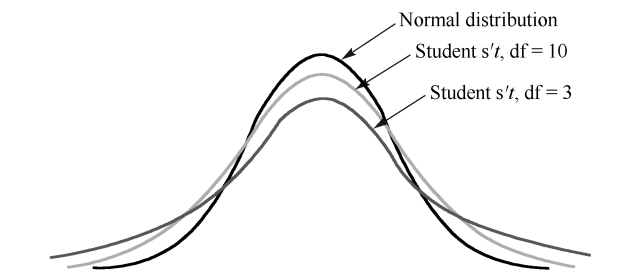


Figure 9.2 Student's t -Distributions

Properties of the t -Distribution ($df > 2$)^①

- (1) t is distributed with a mean of zero.
- (2) t is distributed symmetrically about its mean.
- (3) t is distributed so as to form a family of distributions, a separate distribution for each different number of degrees of freedom ($df \geq 1$):
- (4) The t -distribution approaches the standard normal distribution as the number of degrees of freedom increases.

(5) t is distributed with a variance greater than 1, but as the degrees of freedom increases, the variance approaches 1.

(6) t is distributed so as to be less peaked at the mean and thicker at the tails than is the normal distribution.

Not all of the properties hold for $df = 1$ and $df = 2$. Since we will not encounter situations where $df = 1$ or 2 , these special cases are not discussed further.

Degrees of Freedom, df

A value that identifies each different distribution of Student's t -distribution. For the methods presented in this chapter, the value of df will be the sample size minus 1: $df = n - 1$.

The number of degrees of freedom associated with s^2 is the divisor ($n - 1$) used to calculate the sample variance s^2 [formula (2.5), Unit 2]; that is, $df = n - 1$. The sample variance is the mean of the squared deviations. The number of degrees of freedom is the “number of unrelated deviations” available for use in estimating σ^2 . Recall that the sum of the deviations, $\sum (x - \bar{x})$, must be zero.

From a sample of size n , only the first $n - 1$ of these deviations has freedom of value. That is, the last, or n th, value of $(x - \bar{x})$ must make the sum of the n deviations total exactly zero. As a result, variance is said to average $n - 1$ unrelated squared deviation values, and this number, $n - 1$, was named “degrees of freedom”.

Although there is a separate t -distribution for each degree of freedom, $df = 1$, $df = 2, \dots$, $df = 20, \dots$, $df = 40$, and so on, only certain key critical values of t will be necessary for our work. Consequently, the table for the Student's t -distribution (Statistical Table 4 in Appendix) is a table of critical values rather than a complete table, such as Table 1 is for the standard normal distribution for z . As you look at Table 4, you will note that the left side of the table is identified by “ df ”, degrees of freedom. This left-hand column starts at 3 at the top and lists consecutive df values to 30, then jumps to 35, ..., to “ $df > 100$ ” at the bottom. As we stated, as the degrees of freedom increases, the t -distribution approaches the characteristics of the standard normal z -distribution, see Figure 9.2. Once df is “greater than 100,” the critical values of the t -distribution are the same as the corresponding critical values of the standard normal distribution as given in Statistical Table 2(II) in Appendix.

9.1.1 Using the t -Distribution Table

The critical values of the Student's t -distribution that are to be used both for constructing a confidence interval and for hypothesis testing will be obtained from Statistical Table 4 in Appendix. To find the value of t , you will need to know two identifying values: (1) df , the number of degrees of freedom (identifying the distribution of interest), and (2) α , the area under the curve to the right of the right-hand critical value. A notation much like that used with z will be used to identify a critical value. $t(df, \alpha)$, read as “ t of df , α ,” is the symbol for the value of t with df degrees of freedom and an area of α in the right-hand tail, as shown in Figure 9.3.

Finding t in Relation to the Mean

There are three relationships of t to the mean: t can be on the right side of the mean, on the left side, or have values that bound a certain percentage. Let's start by finding the value of t on the right side of the mean, specifically finding the value of $t(10, 0.05)$, see Figure 9.4.

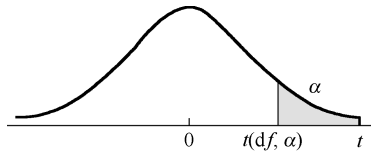


Figure 9.3 t -Distribution Showing $t(df, \alpha)$

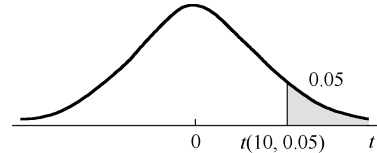


Figure 9.4 Finding the value of $t(10, 0.05)$

Example 9.1 Find the value of $t(10, 0.05)$ (see the diagram 9.4).

There are 10 degrees of freedom, and 0.05 is the area to the right of the critical value. In Table 4 of Appendix Tables, we look for the row $df = 10$ and the column marked “Amount of α in One Tail,” $\alpha = 0.05$. At their intersection, we see that $t(10, 0.05) = 1.81$, see Table 9.1.

Table 9.1 Portion of Statistical Table 4

df	Portion of Statistical Table 4		
	Amount of α in One Tail		
...	0.05	...	
10	1.81		

$$t(10, 0.05) = 1.81$$

For the values of t on the left side of the mean, we can use one of two notations. The t -value shown in Figure 9.6 could be named $t(df, 0.95)$, since the area to the right of it is 0.95, or it could be identified by $-t(df, 0.05)$, since the t -distribution is symmetric about its mean, zero.

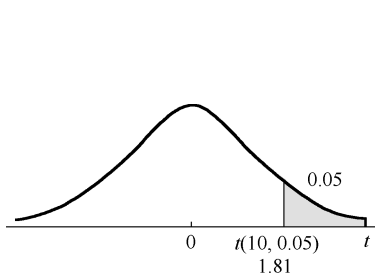


Figure 9.5 Finding the value of $t(10, 0.05)$

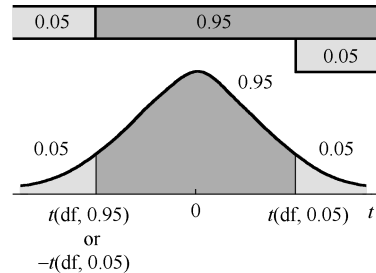


Figure 9.6 t -Value on the left side

Example 9.2 Find the value of $t(15, 0.05)$

There are 15 degrees of freedom. In Statistical Table 4 we look for the column marked $\alpha = 0.05$ (one tail) and its intersection with the row $df = 15$. The table gives us $t(15, 0.05) = 1.75$; therefore, $t(15, 0.95) = -t(15, 0.05) = -1.75$. The value is negative because it is to the left of the mean; see Figure 9.7.

Example 9.3 Finding t -values that bound a middle percentage.

We can find the values of the t -distribution that bound the middle 0.90 of the area under the curve for the distribution with $df = 17$.

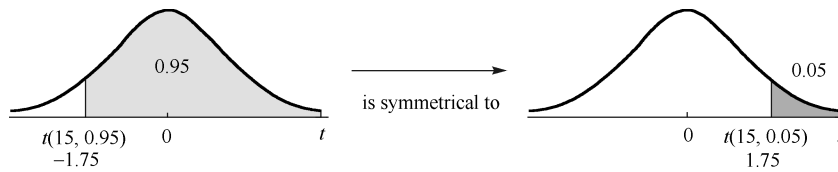


Figure 9.7 The value of $t(15, 0.05)$ and the value of $t(15, 0.95)$

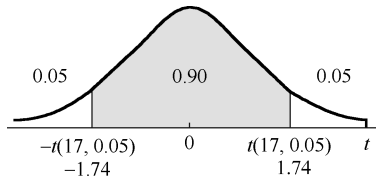


Figure 9.8 The t -distribution is symmetric about its mean

The middle 0.90 leaves 0.05 for the area of each tail. The value of t that bounds the right-hand tail is $t(17, 0.05) = 1.74$, as found in Statistical Table 4. The value that bounds the left-hand tail is **-1.74** because the t -distribution is symmetric about its mean, zero. See Figure 9.8.

If the df needed is not listed in the left-hand column of Table 4, then use the next smaller value of df that is listed. For example, $t(72, 0.05)$ is estimated using $t(70, 0.05) = 1.67$.

9.1.2 Confidence Interval Procedure

We are now ready to make inferences about the population mean μ using the sample standard deviation. As we mentioned earlier, use of the t -distribution has a condition:

The assumption for inferences about the mean μ when σ is unknown:

The sampled population is normally distributed.

The procedure to make confidence intervals using the sample standard deviation is very similar to that used when σ is known (see section 8.2). The difference is the use of the Student's t in place of the standard normal z and the use of s , the sample standard deviation, as an estimate of σ . The central limit theorem implies that this technique can also be applied to nonnormal populations when the sample size is sufficiently large.

Confidence Interval for Mean:

$$\bar{x} - t(df, \alpha/2) \left(\frac{s}{\sqrt{n}} \right) \text{ to } \bar{x} + t(df, \alpha/2) \left(\frac{s}{\sqrt{n}} \right),$$

$$\text{with } df = n - 1 \quad (9.1)$$

Confidence Interval for μ with σ Unknown

Example 9.4

To illustrate how confidence intervals can be formed utilizing the t -distribution, consider a random sample of 20 weights taken from babies born at Northside Hospital. A mean of 6.87 lb and a standard deviation of 1.76 lb were found for the sample. Based on past information, it is assumed that weights of newborns are normally distributed. Using the five-step process, we can estimate, with 95% confidence, the mean weight of all babies born in this hospital.

Step 1 The Set-Up:

Describe the population parameter of interest.

μ , the mean weight of newborns at Northside Hospital.

Step 2 The Confidence interval criteria:

a. Check the assumptions.

σ is unknown, and past information indicates that the sampled population is normal.

b. Identify the probability distribution and the formula to be used.

The Student's t -distribution will be used with formula (9.1).

c. State the level of confidence: $1-\alpha = 0.95$.

Step 3 The Sample evidence:

Collect the sample information: $n = 20$, $\bar{x} = 6.87$, and $s = 1.76$.

Step 4 The Confidence interval:

a. Determine the confidence coefficients.

Since $1-\alpha = 0.95$, $\alpha = 0.05$: therefore $\alpha/2 = 0.025$. Also, since $n = 20$, $df = 19$. At the intersection of row $df = 19$ and one-tailed column $\alpha = 0.025$ in Statistical Table 4, we find $t(df, \alpha/2) = t(19, 0.025) = 2.09$, see Figure 9.9.

Information about the confidence coefficient and using Statistical Table 4 is in Section 9.1.1.

b. Find the maximum error of estimate.

$$E = t(df, \alpha / 2) \left(\frac{s}{\sqrt{n}} \right):$$

$$E = t(19, 0.025) \left(\frac{s}{\sqrt{n}} \right) = 2.09 \left(\frac{1.76}{\sqrt{20}} \right) = (2.09)(0.394) = 0.82$$

c. Find the lower and upper confidence limits.

$$\bar{x} - E \text{ to } \bar{x} + E$$

$$6.87 - 0.82 \text{ to } 6.87 + 0.82$$

$$6.05 \text{ to } 7.69$$

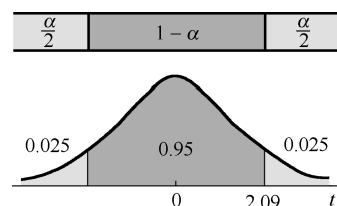


Figure 9.9 Finding the value of $t(19, 0.025)$

Step 5 The Results: State the confidence interval.

6.05 to 7.69, the 95% confidence interval for μ . That is, with 95% confidence we estimate the mean weight to be between 6.05 and 7.69 lb.

9.1.3 Hypothesis-Testing Procedure

The t -statistic is used to complete a hypothesis test about the population mean μ in much the same manner z was used in Unit 8. In hypothesis-testing situations, we use formula (9.2) to calculate the value of the test statistic t^* :

Test Statistic for Mean:

$$t^* = \frac{\bar{x} - \mu}{s / \sqrt{n}}, \text{ with } df = n-1 \quad (9.2)$$

The calculated t is the number of estimated standard errors \bar{x} is from the hypothesized mean

μ . As with confidence intervals, the central limit theorem indicates that the t -distribution can also be applied to nonnormal populations when the **sample size** is sufficiently large.

One-Tailed Hypothesis Test for μ with σ Unknown

Example 9.5

To conduct a one-tailed hypothesis test for μ with σ unknown, let's return to the hypothesis of the example from Unit 8 (see section 8.4) where the EPA wanted to show that the mean carbon monoxide level is higher than 4.9 parts per million. Does a random sample of 22 readings (sample results: $\bar{x} = 5.1$ and $s = 1.17$) present sufficient evidence to support the EPA's claim? Use $\alpha = 0.05$. Previous studies have indicated that such readings have an approximately normal distribution. Again we'll follow the five-step procedure:

Step 1 The Set-Up:

- Describe the population parameter of interest.
 μ , the mean carbon monoxide level of air in downtown Rochester.
- State the null hypothesis (H_o) and the alternative hypothesis (H_a).

$$H_o: \mu = 4.9 (\leq) \text{ (no higher than)}$$

$$H_a: \mu > 4.9 \text{ (higher than)}$$

Step 2 The Hypothesis Criteria:

a. Check the assumptions.

The assumptions are satisfied because the sampled population is approximately normal and the sample size is large enough for the CLT to apply; σ is unknown.

b. Identify the probability distribution and the test statistic to be used.

The t -distribution with $df = n - 1 = 21$, and the test statistic is t^* , formula (9.2).

c. Determine the level of significance: $\alpha = 0.05$.

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 22$, $\bar{x} = 5.1$, and $s = 1.17$.

b. Calculate the value of the test statistic.

Use formula (9.2):

$$t^* = \frac{\bar{x} - \mu}{s / \sqrt{n}} : t^* = \frac{5.1 - 4.9}{1.17 / \sqrt{22}} = \frac{0.20}{0.2494} = 0.8018 \approx 0.80$$

Step 4 The Probability Distribution:

As always, we can use either the p -value procedure or the classical procedure.

(i) Using the p -value procedure:

- Calculate the p -value for the test statistic. Use the right-hand tail because H_a expresses concern for values related to "higher than". $P = P(t^* > 0.80, \text{ with } df = 21)$ as shown in Figure 9.10.

To find the p -value, use one of three methods:

- Use Statistical Table 4 in Appendix to place bounds on the p -value: $0.10 < P < 0.25$.
- Use Statistical Table 5 in Appendix to read the value directly: $P = 0.216$.

3. Use a computer or calculator to calculate the p -value: $P = 0.2163$.
- b. Determine whether or not the p -value is smaller than α .

The p -value is not smaller than α , the level of significance.

(ii) Using the classical procedure:

- a. Determine the critical region and critical value(s).

The critical region is the right-hand tail because H_a expresses concern for values related to “higher than”. The critical value is found at the intersection of the $df = 21$ row and the onetailed 0.05 column of Table 6: $t(21, 0.05) = 1.72$, see Figure 9.10.

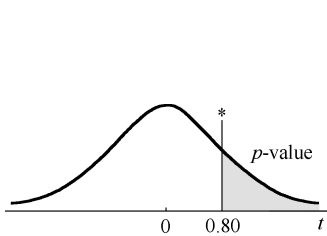


Figure 9.10 Finding p -value

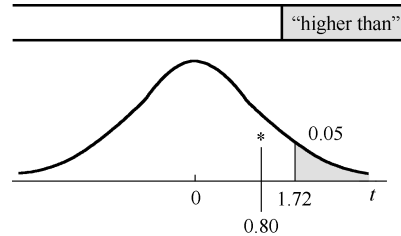


Figure 9.11 Finding p -value using the classical procedure

- b. Determine whether or not the calculated test statistic is in the critical region.
- t^* is not in the critical region, as shown in red in the figure above.

Step 5 The Results:

- a. State the decision about H_o : Fail to reject H_o .
- b. State the conclusion about H_a .

At the 0.05 level of significance, the EPA does not have sufficient evidence to show that the mean carbon monoxide level is higher than 4.9.

Calculating the p -value when using the t -distribution

Method 1: Use Statistical Table 6 in Appendix to place bounds on the p -value. By inspecting the $df=21$ row of Statistical Table 6, you can determine an interval within which the p -value lies. Locate t^* along the row labeled $df=21$. If the t^* value is not listed, locate the two table values it falls between and read the bounds for the p -value from the top of the table. In this case, $t^* = 0.80$ is between 0.686 and 1.32; therefore, P is between 0.10 and 0.25. Use the one-tailed heading, since H_a is one-tailed in this illustration. (Use the two-tailed heading when H_a is two-tailed.)

The 0.686 entry in the table tells us that $P(t > 0.686) = 0.25$, as shown in purple on the figure below. The 1.32 entry in the table tells us that $P(t > 1.32) = 0.10$, as shown in green. You can see that the p -value P (shown in black) is between 0.10 and 0.25, see Figure 9.12. Therefore, $0.10 < P < 0.25$, and we say that 0.10 and 0.25 are the “bounds” for the p -value.

Method 2: Use Statistical Table 7 in Appendix to read the p -value or to “place bounds” on the p -value. Statistical Table 7 is designed to yield p -values given the t^* and df values or to produce bounds on P that are narrower than those produced by Statistical Table 6.

Finding $P = P(t^* > 0.80, \text{ with } df = 21)$

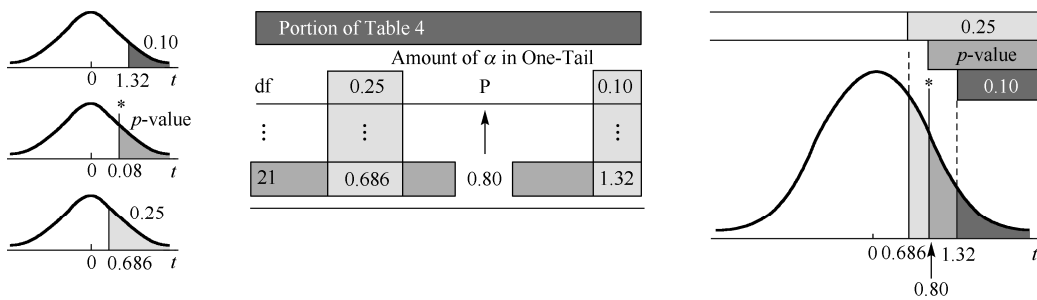


Figure 9.12 Finding p -value using the t -distribution

Table 9.2 Portion of Statistical Table 5

Portion of Statistical Table 5		
df	...	21
t^*		
\vdots		
0.80		0.216

$P = P(t^* > 0.80, \text{ with } df=21)$
=0.216

In the preceding example, $t^* = 0.80$ and $df = 21$. These happen to be row and column headings, so the p -value can be read directly from the table. Locate the p -value at the intersection of the $t^* = 0.80$ row and the $df = 21$ column. The p -value for $t^* = 0.80$ with $df = 21$ is 0.216, see Table 9.2.

To illustrate how to place bounds on the p -value when t^* and df are not the heading values, let's consider the situation where $t^* = 2.43$ with $df = 16$. The $t^* = 2.43$ is between rows $t = 2.4$ and $t = 2.5$, while $df = 16$ is between columns $df = 15$ and $df = 18$. These two rows and two columns intersect a total of four times, namely at 0.015 and 0.014 in the row $t^* = 2.4$ and at 0.012 and 0.011 in the row $t^* = 2.5$. The p -value we are looking for is bounded by the smallest and largest of these four values, namely, 0.011 (lower right) and 0.015 (upper left). Therefore, the bounds for the p -value are $0.011 < P < 0.015$, see Table 9.3.

Table 9.3 Portion of Statistical Table 5

Portion of Statistical Table 5					
t^*	df	...	15	16	18
\vdots					
2.4			0.015		0.014
2.43				P	
2.5			0.012		0.011

$P = P(t^* > 2.43, \text{ with } df=16)$
 $0.011 < P < 0.015$

Method 3: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value for you.

Two-Tailed Hypothesis Test For μ with σ Unknown

Let's look at a two-tailed hypothesis-testing situation, which we can also do for μ with σ unknown.

Example 9.6

This time, we'll examine data from a popular self-image test that results in normally distributed scores. The mean score for public-assistance recipients is expected to be 65. A random

sample of 28 public-assistance recipients in Emerson County is given the test. They achieve a mean score of 62.1, and their scores have a standard deviation of 5.83. Do the ABC Company public-assistance recipients test differently, on the average, than what is expected, at the 0.02 level of significance? To find out, we again turn to our five-step procedure.

Step 1 The Set-Up:

a. Describe the population parameter of interest.

μ , the mean self-image test score for all ABC Company public-assistance recipients.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

H_o : $\mu = 65$ (mean is 65)

H_a : $\mu \neq 65$ (mean is different from 65)

Step 2 The Hypothesis Test Criteria:

a. Check the assumptions.

The test is expected to produce normally distributed scores; therefore, the assumption has been satisfied; σ is unknown.

b. Identify the probability distribution and the test statistic to be used.

The t -distribution with $df = n - 1 = 27$, and the test statistic is t^* , formula (9.2).

c. Determine the level of significance: $\alpha = 0.02$ (given in statement of problem).

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 28$, $\bar{x} = 62.1$, and $s = 5.83$.

b. Calculate the value of the test statistic.

Use formula (9.2):

$$t^* = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

$$t^* = \frac{62.1 - 65.0}{5.83 / \sqrt{28}} = \frac{-2.9}{1.118} = -2.632 \approx -2.63$$

Step 4 The Probability Distribution:

Again, we can choose either the p -value or classical procedure.

(i) Using the p -value procedure:

a. Calculate the p -value for the test statistic. Use both tails because H_a expresses concern for values related to “different from”. $P = P(t < -2.63) + P(t > 2.63) = 2 \cdot P(t > 2.63)$, with $df = 27$ as shown in Figure 9.13.

To find the p -value, use one of three methods:

1. Use Statistical Table 4 in Appendix to place bounds on the p -value: $0.01 < P < 0.02$.
2. Use Statistical Table 5 in Appendix to place bounds on the p -value: $0.012 < P < 0.016$.
3. Use a computer or calculator to calculate the p -value: $P = 0.0140$.

Specific details follow this example.

b. Determine whether or not the p -value is smaller than α .

The p -value is smaller than the level of significance, α .

(ii) Using the classical procedure:

a. Determine the critical region and critical value(s).

The critical region is both tails because H_a expresses concern for values related to “different from”. The critical value is found at the intersection of the $df = 27$ row and the one-tailed 0.01 column of Statistical Table 4: $t(27, 0.01) = 2.47$.

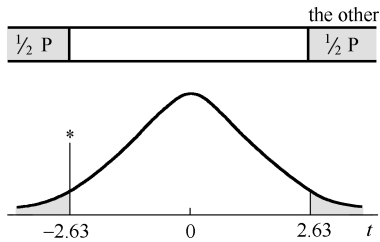


Figure 9.13 Finding p -value

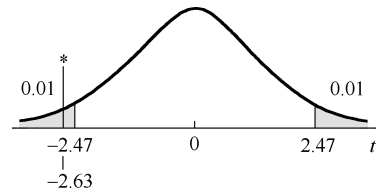


Figure 9.14 Finding p -value using the classical procedure

b. Determine whether or not the calculated test statistic is in the critical region. t^* is in the critical region, as shown in red in the preceding figure.

Step 5 The Results:

a. State the decision about H_o : Reject H_o .

b. State the conclusion about H_a .

At the 0.02 level of significance, we do have sufficient evidence to conclude that the ABC Company assistance recipients' test results are significantly different, on the average, from the expected 65.

Calculating the p-value when using the t-distribution

Method 1: Using Statistical Table 4, find 2.63 between two entries in the $df = 27$ row and read the bounds for P from the two-tailed heading at the top of the table:

$$0.01 < P < 0.02.$$

Method 2: Generally, bounds found using Table 5 will be narrower than bounds found using Table 4. The table at right shows you how to read the bounds from Table 5; find $t^* = 2.63$ between two rows and $df = 27$ between two columns, and locate the four intersections of these columns and rows. The value of $\frac{1}{2}P$ is bounded by the upper left and the lower right of these table entries, see Table 9.4.

Table 9.4 Portion of Statistical Table 5

Portion of Statistical Table 5			
Degrees of Freedom			
t^*	.25	27	29
\vdots			\vdots
2.6	0.008		0.007
2.63		$\frac{1}{2}P$	
2.7	0.006		0.006

$P = 2P(T^* > 2.63, \text{ with } df=27)$

$0.006 < \frac{1}{2}P < 0.008$

$0.012 < \frac{1}{2}P < 0.016$

Method 3: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value for you (do not double it).

New Words and Expressions

Irish ['aɪrɪʃ] *n.* 爱尔兰人；爱尔兰语 *adj.* 爱尔兰的，爱尔兰人的
brewery ['bru:əri] *n.* 啤酒厂，酿酒厂
pseudonym ['su:dənɪm] *n.* (尤指) 笔名，假名，化名
diagrammatic [ˌdaɪəgrə'mætɪk] *adj.* 图表的，概略的
self-image [self'ɪmɪdʒ] *n.* 自我形象，自我印象
public-assistance 政府资助；社会援助；公共救助
recipient [rɪ'sɪpiənt] *n.* 接受者；容器；容纳者 *adj.* 容易接受的；感受性强的

Technical Terms

degrees of freedom 自由度

9.2 Inferences about the Binomial Probability of Success

Perhaps the most common inference involves the *binomial parameter* p , the “probability of success”.

Yes, every one of us uses this inference, even if only casually. In thousands of situations we are concerned about something either “happening” or “not happening”. There are only two possible outcomes of concern, and that is the fundamental property of a **binomial experiment**. The other necessary ingredient is multiple independent trials. Asking five people whether they are “for” or “against” some issue can create five independent trials; if 200 people are asked the same question, 200 independent trials may be involved; if 30 items are inspected to see if each “exhibits a particular property” or “not” there will be 30 repeated trials; these are the makings of a binomial inference.

The binomial parameter p is defined to be the probability of success on a single trial in a binomial experiment.

Sample Binomial Probability

$$p' = \frac{x}{n} \quad (9.3)$$

where the **random variable** x represents the number of successes that occur in a sample consisting of n trials. Recall that the mean and standard deviation of the binomial random variable x are found by using formula (5.7), $\mu = np$, and formula (5.8), $\sigma = \sqrt{npq}$, where $q = 1 - p$. The distribution of x is considered to be approximately normal if n is greater than 20 and if np and nq are both greater than 5. This commonly accepted *rule of thumb* allows us to use the **standard normal distribution** to estimate probabilities for the binomial random variable x , the number of successes

in n trials, and to make inferences concerning the binomial parameter p , the probability of success on an individual trial.

Generally, it is easier and more meaningful to work with the distribution of p' (the observed probability of occurrence) than with x (the number of occurrences). Consequently, we will convert formulas (5.7) and (5.8) from units of x (integers) to units of proportions (percentages expressed as decimals) by dividing each formula by n , as shown in Figure 9.15.

The information about the sampling distribution of p' is summarized as follows:

If a random sample of size n is selected from p large population with $p = P(\text{success})$, then the sampling distribution of p' has:

- (i) A mean $\mu_{p'}$, equal to p ,
- (ii) A standard error $\sigma_{p'}$, equal to $\sqrt{\frac{pq}{n}}$,
- (iii) An approximately normal distribution if n is sufficiently large .

In practice, using these guidelines will ensure normality:

1. The sample size is greater than 20.
2. The products np and nq are both greater than 5.
3. The sample consists of less than 10% of the population.

We are now ready to make inferences about the population parameter p . Use of the z -distribution involves an assumption:

The assumption for inferences about the binomial parameter p :

The n random observations that form the sample are selected independently from a population that is not changing during the sampling.

9.2.1 Confidence Interval Procedure

Inferences concerning the population binomial parameter p , $P(\text{success})$, are made using procedures that closely parallel the inference procedures used for the population mean μ . When we estimate the **population proportion** p , we will base our estimations on the **point estimate** p' , as shown in Figure 9.15. The point estimate, p' , becomes the center of the confidence interval, and the maximum error of estimate is a multiple of the **standard error**. The **level of confidence** determines the confidence coefficient, the number of multiples of the standard error.

Variable	Mean	Standard Deviation
x	$\mu_x = np$	$\sigma_x = \sqrt{npq}$
to change x to p' , divide by n	$\frac{np}{n}$	$\frac{\sqrt{npq}}{n}$
p'	$\mu_{p'} = p \quad (9.4)$	$\sigma_{p'} = \sqrt{\frac{pq}{n}} \quad (9.5)$

Figure 9.15 Formulas (9.4) and (9.5)

Confidence Interval for a Proportion

$$p' - z(\alpha / 2) \left(\frac{\sqrt{p'q'}}{n} \right) \text{ to } p' + z(\alpha / 2) \left(\frac{\sqrt{p'q'}}{n} \right) \quad (9.4)$$

where $p' = \frac{x}{m}$ and $q' = 1 - p'$.

Notice that the standard error, $\sqrt{\frac{pq}{n}}$, has been replaced by $\sqrt{\frac{p'q'}{n}}$. Since we are estimating p , we do not know its value and therefore we must use the best replacement available. That replacement is p' , the observed value or the point estimate for p . This replacement will cause little change in the standard error or the width of our confidence interval provided n is sufficiently large.

Confidence Interval for p

We can illustrate the formation of a confidence interval for the binomial parameter p with the following example.

Example 9.7

In a discussion about the cars that fellow students drive, several statements were made about types, ages, makes, colors, and so on. Dana decided he wanted to estimate the proportion of convertibles students drive, so he randomly identified 200 cars in the student parking lot and found 17 to be convertibles. To find the 90% confidence interval for the proportion of convertibles driven by students, we once again follow the five-step process.

Step 1 The Set-Up:

Describe the population parameter of interest.

p , the proportion (percentage) of convertibles driven by students.

Step 2 The Confidence Interval Criteria:

a. Check the assumptions.

The sample was randomly selected, and each student's response is independent of those of the others surveyed.

b. Identify the probability distribution and the formula to be used.

The standard normal distribution will be used with formula (9.6) as the test statistic, p' is expected to be approximately normal because: (1) $n = 200$ is greater than 20, and (2) both np [approximated by $np' = 200(17/200) = 17$] and nq [approximated by $nq' = 200(183/200) = 183$] are greater than 5.

c. State the level of confidence: $1 - \alpha = 0.90$.

Step 3 The Sample Evidence:

Collect the sample information. $n = 200$ cars were identified, and $x = 17$ were convertibles:

$$p' = \frac{x}{n} = \frac{17}{200} = 0.085$$

Step 4 The Confidence Interval:

a. Determine the confidence coefficient.

This is the z -score [$z(\alpha/2)$, "z of one-half of alpha"] identifying the number of standard errors

needed to attain the level of confidence and is found using Statistical Table 2 in Appendix; $z(\alpha/2) = z(0.05) = 1.65$, see Figure 9.16.

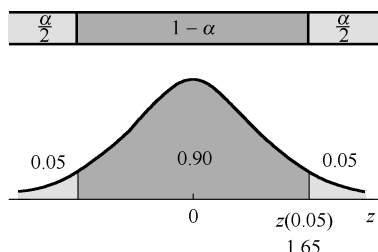


Figure 9.16 Finding the value of $z(0.05)$

b. Find the maximum error of estimate. Use the maximum error part of formula (9.6):

$$E = z(\alpha/2) \sqrt{\frac{p'q'}{n}} = 1.65 \left(\sqrt{\frac{(0.085)(0.915)}{200}} \right) \\ = (1.65) \sqrt{0.000389} = (1.65)(0.020) = 0.033$$

c. Find the lower and upper confidence limits.

$$\begin{array}{lll} p' - E & \text{to} & p' + E \\ 0.085 - 0.033 & \text{to} & 0.085 + 0.033 \\ 0.052 & \text{to} & 0.118 \end{array}$$

Step 5 The Results:

State the confidence interval.

0.052 to 0.118 is the 90% confidence interval for $p = P(\text{drives convertible})$.

That is, the true proportion of students who drive convertibles is between 0.052 and 0.118, with 90% confidence.

9.2.2 Determining Sample Size

By using the maximum error part of the confidence interval formula, it is possible to determine the **size of the sample** that must be taken in order to estimate p with a desired accuracy. Here is the formula for the **maximum error of estimate for a proportion**:

$$E = z(\alpha/2) \left(\sqrt{\frac{pq}{n}} \right) \quad (9.5)$$

To determine the sample size from this formula, we must decide on the quality we want for our final confidence interval. This quality is measured in two ways: the level of confidence and the preciseness (narrowness) of the interval. The level of confidence we establish will in turn determine the confidence coefficient, $z(\alpha/2)$. The desired preciseness will determine the maximum error of estimate, E . (Remember that we are estimating p , the binomial probability; therefore, E will typically be expressed in hundredths.)

For ease of use, we can solve formula (9.7) for n as follows:

Sample Size for $1 - \alpha$ Confidence Interval for p

$$n = \frac{[z(\alpha/2)]^2 \cdot p^* \cdot q^*}{E^2} \quad (9.6)$$

where p^* and q^* are provisional values of p and q used for planning.

By inspecting formula (9.8), we can observe that three components determine the sample size:

- (1) The level of confidence $[1 - \alpha]$, which determines the confidence coefficient, $z(\alpha/2)$;
- (2) The provisional value of p (p^* determines the value of q^*);
- (3) The maximum error, E .

An increase or decrease in one of the three components shown in Figure 9.17 affects the sample size. If the level of confidence is increased or decreased (while the other components are held constant), then the sample size will increase or decrease, respectively. If the product of p^* and q^* is increased or decreased (with other components held constant), then the sample size will increase or decrease, respectively. (The product $p^* \cdot q^*$ is largest when $p^* = 0.5$ and decreases as the value of p^* moves farther from 0.5.) An increase or decrease in the desired maximum error will have the opposite effect on the sample size, since E appears in the denominator of the formula. If no provisional values for p and q are available, then use $p^* = 0.5$ and $q^* = 0.5$. Using $p = 0.5$ is safe because it gives the largest sample size of any possible value of p . Using $p^* = 0.5$ works reasonably well when the true value is “near 0.5” (say, between 0.3 and 0.7); however, as p gets nearer to either zero or one, a sizable overestimate in sample size will occur.

Sample Size for Estimating p (No Prior Information)

Example 9.8

Using confidence intervals, we can determine the sample size required for estimating p with no prior information. For example, to find the sample size required to estimate the true proportion of blue-eyed community college students if you want your estimate to be within 0.02 with 90% confidence, we would do the following:

Step 1 The level of confidence is $1 - \alpha = 0.90$; therefore, the confidence coefficient is $z(\alpha/2) = z(0.05) = 1.65$ from Statistical Table 2 in Appendix; see Figure 9.16.

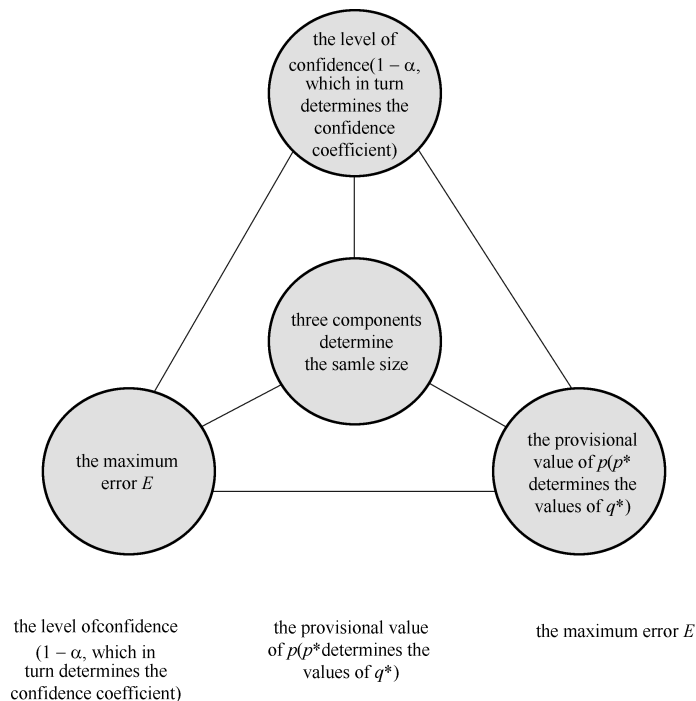


Figure 9.17 The sample size and three components

Step 2 The desired maximum error is $E = 0.02$.

Step 3 Since no estimate was given for p , use $p^* = 0.5$ and $q^* = 1 - p^* = 0.5$.

Step 4 Use formula (9.8) to find n :

$$n = \frac{[z(\alpha/2)]^2 \cdot p^* \cdot q^*}{E^2}$$
$$n = \frac{(1.65)^2 \cdot 0.5 \cdot 0.5}{(0.02)^2} = \frac{0.680625}{0.0004} = 1701.56 \approx 1702$$

Note: When finding the sample size n , always round up to the next larger integer, no matter how small the decimal.

Sample Size for Estimating p (Prior Information)

Example 9.9

We can also determine the sample size for estimating p when we do have prior information. Consider an automobile manufacturer that purchases bolts from a supplier who claims the bolts are approximately 5% defective. To determine the sample size required to estimate the true proportion of defective bolts if we want our estimate to be within ± 0.02 with 90% confidence, we would do the following:

Step 1 The level of confidence is $1 - \alpha = 0.90$; the confidence coefficient is $z(\alpha/2) = z(0.05) = 1.65$.

Step 2 The desired maximum error is $E = 0.02$.

Step 3 Since there is an estimate for p (supplier's claim is "5% defective"), use $p^* = 0.05$ and $q^* = 1 - p^* = 0.95$.

Step 4 Use formula (9.8) to find n :

$$n = \frac{[z(\alpha/2)]^2 \cdot p^* \cdot q^*}{E^2}$$
$$n = \frac{(1.65)^2 \cdot 0.05 \cdot 0.95}{(0.02)^2} = 0.12931875 / 0.0004 = 323.3 \approx 324$$

Notice the difference in the sample sizes required in the two previous examples (with and without prior information). The only mathematical difference between the problems is the value used for p^* . In the first example, we used $p^* = 0.5$, and in the second example we used $p^* = 0.05$. Recall that the use of the provisional value $p^* = 0.5$ gives the maximum sample size. As you can see, it will be an advantage to have some indication of the value expected for p , especially as p moves increasingly farther from 0.5.

9.2.3 Hypothesis-Testing Procedure

When the binomial parameter p is to be tested using a hypothesis-testing procedure, we will use a test statistic that represents the difference between the observed proportion and the hypothesized proportion, divided by the standard error. This test statistic is assumed to be normally distributed when the null hypothesis is true, when the assumptions for the test have been satisfied, and when n is sufficiently large ($n > 20$, $np > 5$, and $nq > 5$).

Test Statistic for a Proportion

$$z^* = \frac{p' - p}{\sqrt{\frac{pq}{n}}} \quad \text{with} \quad p' = \frac{x}{n} \quad (9.7)$$

To demonstrate the use of this formula, we'll use two examples: one- and two-tailed hypothesis tests for proportion p .

One-Tailed Hypothesis Test for Proportion p

Example 9.10

Many people sleep in on the weekends to make up for “short nights” during the workweek. The Better Sleep Council reports that 61% of us get more than seven hours of sleep per night on the weekend. A random sample of 350 adults found that 235 had more than seven hours of sleep each night last weekend. At the 0.05 level of significance, does this evidence show that more than 61% sleep seven or more hours per night on the weekend?

Step 1 The Set-Up:

a. Describe the population parameter of interest.

p , the proportion of adults who get more than seven hours of sleep per night on weekends.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

$$H_o: p = P(7+ \text{ hours of sleep}) = 0.61 (\leq) \\ \text{(no more than 61\%)}$$

$$H_a: p > 0.61 \text{ (more than 61\%)}$$

Step 2 The Hypothesis Test Criteria:

a. Check the assumptions.

The random sample of 350 adults was independently surveyed.

b. Identify the probability distribution and the test statistic to be used.

The standard normal z will be used with formula (9.9). Since $n = 350$ is greater than 20 and both $np = (350)(0.61) = 213.5$ and $nq = (350)(0.39) = 136.5$ are greater than 5, p' is expected to be approximately normally distributed.

c. Determine the level of significance: $\alpha = 0.05$.

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 350$ and $x = 235$:

$$p' = \frac{x}{n} = \frac{235}{350} = 0.671$$

b. Calculate the value of the test statistic.

Use formula (9.9):

$$z^* = \frac{p' - p}{\sqrt{\frac{pq}{n}}} : \\ z^* = \frac{0.671 - 0.61}{\sqrt{\frac{(0.61)(0.39)}{350}}} = \frac{0.061}{\sqrt{0.0006797}} = \frac{0.061}{0.0261} = 2.34$$

Step 4 The robability Disdribution:

Again, we can choose either the p -value or classical procedure.

(i) Using the p -value procedure:

a. Calculate the p -value for the test statistic. Use the right-hand tail because H_a expresses concern for values related to “more than.” $P = p\text{-value} = P(z > 2.34)$, as shown in Figure 9.18.

To find the p -value, use one of three methods:

1. Use Statistical Table 1 in Appendix to calculate the p -value:

$$P = 0.5000 - 0.4904 = 0.0096.$$

2. Use Statistical Table 3 in Appendix to place bounds on the p -value:

$$0.0094 < P < 0.0107.$$

3. Use a computer or calculator to calculate the p -value:

$$P = 0.0096.$$

For specific instructions, see *Method 3* at the end of Step 5.

b. Determine whether or not the p -value is smaller than α .

The p -value is smaller than α .

(ii) Using the classical procedure:

a. Determine the critical region and critical value(s).

The critical region is the right-hand tail because H_a expresses concern for values related to “more than”. The critical value is obtained from Statistical Table 2(I): $z(0.05) = 1.65$, see Figure 9.19.

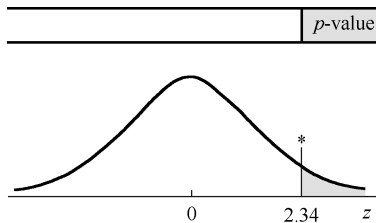


Figure 9.18 Finding p -value

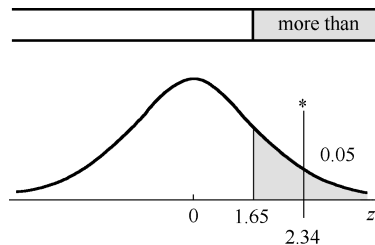


Figure 9.19 Finding p -value using the classical procedure

Specific instructions for finding critical values are given in Section 8.5.

b. Determine whether or not the calculated test statistic is in the critical region.

z^* is in the critical region, as shown in red on the figure above.

Step 4 The Results:

a. State the decision about H_o : Reject H_o .

b. State the conclusion about H_a .

There is sufficient reason to conclude that the proportion of adults in the sampled population who are getting more than seven hours of sleep nightly on weekends is significantly higher than 61% at the 0.05 level of significance.

Method 3: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value for you.

Two-Tailed Hypothesis Test for Proportion p

Example 9.11

Now let's work through a two-tailed hypothesis test for proportion p by picking up the example on page 193 about cars students drive. While talking about the cars that fellow students drive, Tom claimed that 15 % of the students drive convertibles. Jody finds this hard to believe, and she wants to check the validity of Tom's claim using Dana's random sample. At a level of significance of 0.10, we want to determine if there is sufficient evidence to reject Tom's claim if there were 17 convertibles in his sample of 200 cars.

Step 1 The Set-Up:

a. Describe the population parameter of interest.

$p = P(\text{student drives convertible})$.

b. State the null hypothesis (H_0) and the alternative hypothesis (H_a).

$H_0: p = 0.15$ (15% do drive convertibles)

$H_a: p \neq 0.15$ (the percentage is different from 15%)

Step 2 The hypothesis Test Criteria:

a. Check the assumptions.

The sample was randomly selected, and each subject's response is independent of other responses.

b. Identify the probability distribution and the test statistic to be used.

The standard normal z and formula (9.9) will be used. Since $n = 200$ is greater than 20 and both np and nq are greater than 5, p' is expected to be approximately normally distributed.

c. Determine the level of significance: $\alpha = 0.10$.

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 200$ and $x = 17$:

$$p' = \frac{x}{n} = \frac{17}{200} = 0.085$$

b. Calculate the value of the test statistic.

Use formula (9.9):

$$z^* = \frac{p' - p}{\sqrt{\frac{pq}{n}}}$$
$$z^* = \frac{0.085 - 0.150}{\sqrt{\frac{(0.15)(0.85)}{200}}} = \frac{-0.065}{\sqrt{0.00064}} = \frac{-0.065}{0.02525} = -2.57$$

Step 4 The Probability Distribution:

Again, we can choose either the p -value or classical procedure.

(i) Using the p -value procedure:

a. Calculate the p -value for the test statistic.

Use both tails because H_a expresses concern for values related to "different from".

$$P = p\text{-value} = P(z < -2.57) + P(z > 2.57) = 2 \times P(|z| > 2.57)$$

as shown in Figure 9.20.

To find the p -value, use one of three methods:

1. Use Statistical Table 1 in Appendix to calculate the p -value:

$$P = 2 \times (0.5000 - 0.4949) = 0.0102.$$

2. Use Statistical Table 3 in Appendix to place bounds on the p -value:

$$0.0094 < P < 0.0108.$$

3. Use a computer or calculator to calculate the p -value:

$$P = 0.0102.$$

For specific instructions, see pages section 8.4.

b. Determine whether or not the p -value is smaller than α .

The p -value is smaller than α .

(ii) Using the classical procedure:

a. Determine the critical region and critical value(s).

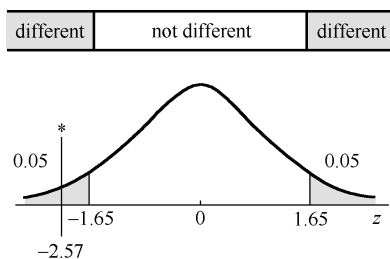


Figure 9.21 Finding p -value using the classical procedure

The critical region is two-tailed because H_a expresses concern for values related to “different from”. The critical value is obtained from Statistical Table 2(II): $z(0.05) = 1.65$, see Figure 9.21.

For specific instructions, see Section 8.5.

b. Determine whether or not the calculated test statistic is in the critical region.

z^* is in the critical region, as shown in red on the figure above.

Step 4 The Results:

a. State the decision about H_0 : Reject H_0 .

b. State the conclusion about H_a .

There is sufficient evidence to reject Tom’s claim and conclude that the percentage of students who drive convertibles is different from 15% at the 0.10 level of significance.

New Words and Expressions

replacement [rɪˈpleɪsmənt] *n.* 代替；归还，复位；替代者；补充兵员

parking [ˈpɑːkɪŋ] *n.* (车辆等的)停放；停车场所，停车位 *v.* 停车 (park 的 ing 形式)

convertible [kənˈvɜːtəbl] *adj.* 可改变的；可变换的；(汽车等)有折篷的 *n.* 敞篷车

sleep [sliːp] *n.* 睡眠 *vt.* 为……提供床位；提供住宿；以睡觉打发日子

hundredth [ˈhʌndrədθ] *adj.* 第一百的；第一百个的

num. 第 100 个；百分之一；第一百号

Technical Terms

rule of thumb 经验法则, 经验规则

9.3 Inferences about the Variance and Standard Deviation

Problems often arise that require us to make inferences about variability.

For example, a soft-drink bottling company has a machine that fills 16-oz bottles. The company needs to control the standard deviation σ (or variance σ^2) in the amount of soft drink, x , put into each bottle. The mean amount placed in each bottle is important, but a correct mean amount does not ensure that the filling machine is working correctly. If the variance is too large, many bottles will be overfilled and many underfilled. Thus, the bottling company wants to maintain as small a standard deviation (or variance) as possible.

When discussing inferences about the spread of data, we usually talk about variance instead of standard deviation because the techniques (the formulas used) employ the sample variance rather than the standard deviation. However, remember that the standard deviation is the positive square root of the variance; thus, talking about the variance of a population is comparable to talking about the standard deviation.

Background

Inferences about the variance of a normally distributed population use the chi-square, χ^2 , distributions (“kisquare”: that’s “ki” as in “kite” and χ is the Greek lowercase letter chi). The chi-square distributions, like Student’s t -distributions, are a family of probability distributions, each one identified by the parameter number of degrees of freedom. In order to use the chi-square distribution, we must be aware of its properties, see Figure 9.22.

Note: When $df > 2$, the mean value of the chi-square distribution is df . The mean is located to the right of the mode (the value where the curve reaches its high point) and just to the right of the median (the value that splits the distribution, 50% on either side). By locating zero at the left extreme and the value of df on your sketch of the χ^2 distribution, you will establish an approximate scale so that other values can be located in their respective positions, see Figure 9.23.

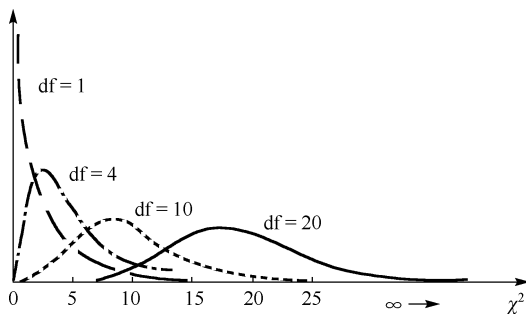


Figure 9.22 Various Chi-Square Distributions

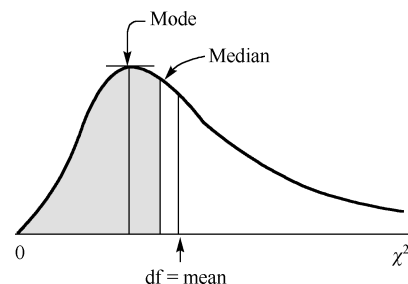


Figure 9.23 Location of Mean, Median, and Mode for χ^2 Distribution

Properties of the Chi-Square Distribution

- (1) χ^2 is nonnegative in value; it is zero or positively valued.
- (2) χ^2 is not symmetrical; it is skewed to the right.
- (3) χ^2 is distributed so as to form a family of distributions, a separate distribution for each different number of degrees of freedom.

9.3.1 Critical Values of Chi-Square

The critical values for chi-square are obtained from Statistical Table 6 in Appendix. Each critical value is identified by two pieces of information: degrees of freedom (df) and area under the

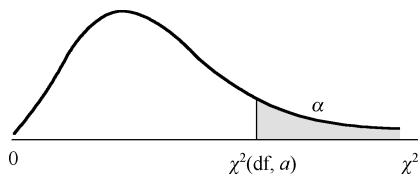


Figure 9.24 Chi-Square Distribution
Showing $\chi^2(df, \alpha)$

curve to the right of the critical value being sought. Thus, $\chi^2(df, \alpha)$ (read “chi-square of *df*, *alpha*”) is the symbol used to identify the critical value of chi-square with *df* degrees of freedom and with α area to the right, as shown in Figure 9.24. Since the chi-square distribution is not symmetrical, the critical values associated with the right and left tails are given separately in Statistical Table 6.

To illustrate finding χ^2 associated with the right tail, let’s find $\chi^2(20, 0.05)$. See Figure 25 below. Use Statistical Table 6 in Appendix to find the value of $\chi^2(20, 0.05)$ at the intersection of row *df* = 20 and column $\alpha = 0.05$, as shown in the portion of the Table 9.5, and see Figure 9.25.

Table 9.5 Portion of Statistical Table 6

df	Portion of Statistical Table 6		
	Area to the Right		
...	0.05	...	
20	3.14		$\chi^2(20, 0.05) = 3.14$

We can also find χ^2 associated with the left tail. To do so, let’s find $\chi^2(14, 0.90)$.

We use Statistical Table 6 in Appendix to find the value of $\chi^2(14, 0.90)$ at the intersection of row *df* = 14 and column $\alpha = 0.90$, as shown in the portion of the Table 9.6.

Table 9.6 Portion of Statistical Table 6

df	Portion of Statistical Table 6		
	Area to the Right		
...	0.90	...	
14	7.79		$\chi^2(14, 0.90) = 7.79$

Applying that number to our curve produces the corresponding figure, see Figure 9.26.

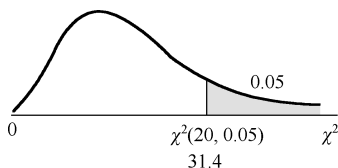


Figure 9.25 Finding $\chi^2(20, 0.05)$ associated with the right tail

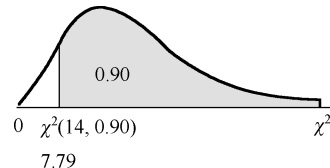


Figure 9.26 $\chi^2(14, 0.90)$ corresponding area

Most computer software packages or statistical calculators will calculate the area related to a specified χ^2 -value. The accompanying figure shows the relationship between the cumulative probability distribution and a specific χ^2 -value for a χ^2 -distribution with df degrees of freedom, see Figure 9.27.

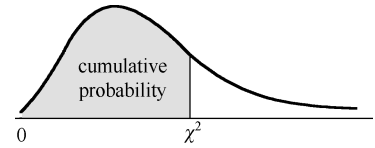


Figure 9.27 χ^2 -value corresponding the area of cumulative probability

9.3.2 Hypothesis-Testing Procedure

We are now ready to use chi-square to make inferences about the population variance or standard deviation.

The assumption for inferences about the variance σ^2 or standard deviation σ :

The sampled population is normally distributed.

The t procedures for inferences about the mean (see section 9.1) were based on the assumption of normality, but they are generally useful even when the sampled population is nonnormal, especially for larger samples. However, the same is not true about the inference procedures for the standard deviation. The statistical procedures for the standard deviation are very sensitive to nonnormal distributions (skewness, in particular), and this makes it difficult to determine whether an apparent significant result is the result of the sample evidence or a violation of the assumptions. Therefore, the only inference procedure to be presented here is the hypothesis test for the standard deviation of a normal population.

The **test statistic** that will be used in testing hypotheses about the population variance or standard deviation is obtained by using the following formula.

Test Statistic for Variance and Standard Deviation

$$\chi^2* = \frac{(n-1)s^2}{\sigma^2}, \quad \text{with df} = n-1 \quad (9.8)$$

When random samples are drawn from a normal population with a known variance σ^2 , the quantity $\frac{(n-1)s^2}{\sigma^2}$ possesses a probability distribution that is known as the chi-square distribution with $n-1$ degrees of freedom.

One-Tailed Hypothesis Test for Variance σ^2

Example 9.12

Let's return to the illustration about the bottling company that wishes to detect when the variability in the amount of soft drink placed into each bottle gets out of control. A variance of 0.0004 is considered acceptable, and the company wants to adjust the bottle-filling machine when the variance, σ^2 , becomes larger than this value. The decision will be made using the hypothesis testing procedure. In this scenario, we're going to conduct a one-tailed hypothesis test for variance, σ^2 .

The soft-drink bottling company wants to control the variability in the amount of fill by not allowing the variance to exceed 0.0004. We need to know if a sample of size 28 with a variance of 0.0007 indicates

that the bottling process is out of control (with regard to variance) at the 0.05 level of significance.

Step 1 The Set-Up:

a. Describe the population parameter of interest.

σ^2 , the variance in the amount of fill of a soft drink during a bottling process.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

$H_o: \sigma^2 = 0.0004$ (\leq) (variance is not larger than 0.0004)

$H_a: \sigma^2 > 0.0004$ (variance is larger than 0.0004)

Step 2 The Hypothesis Test criteria:

a. Check the assumptions.

The amount of fill put into a bottle is generally normally distributed. By checking the distribution of the sample, we could verify this.

b. Identify the probability distribution and the test statistic to be used.

The chi-square distribution will be used and formula (9.10), with $df = n-1 = 28-1 = 27$.

c. Determine the level of significance: $\alpha = 0.05$.

Step 3 The Sample evidence:

a. Collect the sample information: $n = 28$ and $s^2 = 0.0007$.

b. Calculate the value of the test statistic.

Use formula (9.10):

$$\begin{aligned}\chi^{2*} &= \frac{(n-1)s^2}{\sigma^2} : \\ \chi^{2*} &= \frac{(28-1)(0.0007)}{0.0004} \\ &= \frac{(27)(0.0007)}{0.0004} \\ &= 47.25\end{aligned}$$

Step 4 The Probability Distribution:

Again, we can choose either the p -value or classical procedure.

(i) Using the p -value procedure:

a. Calculate the p -value for the test statistic.

Use the right-hand tail because H_a expresses concern for values related to “larger than”. $P = P(\chi^{2*} > 47.25, \text{ with } df = 27)$, as shown on Figure 9.28.

To find the p -value, use one of two methods:

1. Use Statistical Table 6 in Appendix to place bounds on the p -value: $0.005 < P < 0.01$.
2. Use a computer or calculator to calculate the p -value: $P = 0.0093$.

Specific instructions follow this five-step procedure.

b. Determine whether or not the p -value is smaller than α .

The p -value is smaller than the level of significance, α (0.05).

(ii) Using the classical procedure:

a. Determine the critical region and critical value(s).

The critical region is the right-hand tail because H_a expresses concern for values related to “larger than”. The critical value is obtained from Statistical Table 6 at the intersection of row $df = 27$ and column $\alpha = 0.05$: $\chi^2(27, 0.05) = 40.1$.

For specific instructions, see Section 9.3.1.

b. Determine whether or not the calculated test statistic is in the critical region.

χ^{2*} is in the critical region, as shown in black on Figure 9.29.

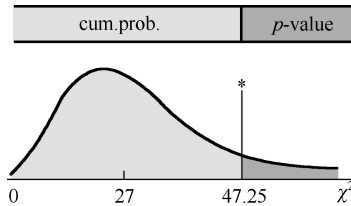


Figure 9.28 Finding p -value

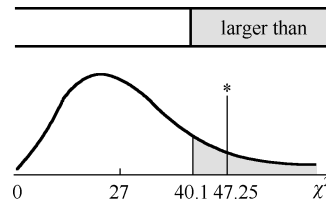


Figure 9.29 Finding p -value using the classical procedure

Step 5 The results:

a. State the decision about H_o : Reject H_o .

b. State the conclusion about H_a .

At the 0.05 level of significance, we conclude that the bottling process is out of control with regard to the variance.

Calculating the p -value when using the χ^2 -distribution

Method 1: Use Statistical Table 6 in Appendix to place bounds on the p -value. By inspecting the $df = 27$ row of Statistical Table 6, you can determine an interval within which the p -value lies. Locate χ^2 , along the row labeled $df = 27$. If χ^2 , is not listed, locate the two values that χ^{2*} falls between, and then read the bounds for the p -value from the top of the table. In this case, $\chi^{2*} = 47.25$ is between 47.0 and 49.6; therefore, P is between 0.005 and 0.01, see Table 9.7.

Table 9.7 Portion of Statistical Table 6

Portion of Statistical Table 6				
df	Area in Right.Hand Tail			
	...	0.01	P	0.005
	2.7	47.0	47.25	49.6

$0.005 < P < 0.01$

Tailed Hypothesis Test for Standard Deviation, σ

Example 9.13

Decisions can also be made using the two-tailed hypothesis-testing procedure. Let’s look at another scenario. The manufacturer of a photographic chemical claims that its product has a shelf life that is normally distributed about a mean of 180 days with a standard deviation of no more than 10 days. As a user of this chemical, Fast Photo is concerned that the standard deviation might be different from 10 days; otherwise, it will buy a larger quantity while the chemical is part of a special promotion. Twelve random samples were selected and tested, with a standard deviation of 14 days resulting. The managers at Fast Photo want to know if, at the 0.05 level of significance, this

sample presents sufficient evidence to show that the standard deviation is different from 10 days.

Step 1 The Set-Up:

a. Describe the population parameter of interest.

σ , the standard deviation for the shelf life of the chemical.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a).

H_o : $\sigma=10$ (standard deviation is 10 days)

H_a : $\sigma \neq 10$ (standard deviation is different from 10 days)

Step 2 The Hypothesis Test Criteria:

a. Check the assumptions.

The manufacturer claims shelf life is normally distributed; this could be verified by checking the distribution of the sample.

b. Identify the probability distribution and the test statistic to be used.

The chi-square distribution will be used and formula (9.10), with $df = n-1 = 12-1 = 11$.

c. Determine the level of significance: $\alpha = 0.05$.

Step 3 The Sample Evidence:

a. Collect the sample information: $n = 12$ and $s = 14$.

b. Calculate the value of the test statistic.

Use formula (9.10):

$$\chi^2* = \frac{(n-1)s^2}{\sigma^2} :$$

$$\chi^2* = \frac{(12-1)(14)^2}{(10)^2} = \frac{2156}{100} = 21.56$$

Step 4 The Probability Distribution:

Again, we can choose either the p -value or classical procedure.

(i) Using the p -value procedure:

a. Calculate the p -value for the test statistic.

Since the concern is for values “different from” 10, the p -value is the area of both tails. The area of each tail will represent $\frac{1}{2}P$. Since $\chi^2* = 21.56$ is in the right tail, the area of the right tail is $\frac{1}{2}P$.

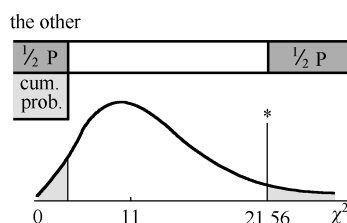


Figure 9.30 Finding p -value

$$\frac{1}{2}P = P(\chi^2 > 21.56, \text{ with } df = 11)$$

as shown in Figure 9.30.

To find $\frac{1}{2}P$, use one of two methods:

1. Use Statistical Table 6 in Appendix to place bounds on

$$\frac{1}{2}P : 0.025 < \frac{1}{2}P < 0.05. \text{ Double both bounds to find the bounds}$$

for P : $2 \times (0.025 < \frac{1}{2}P < 0.05)$ becomes $0.05 < P < 0.10$.

2. Use a computer or calculator to find $\frac{1}{2}P$: $\frac{1}{2}P = 0.0280$; therefore, $P = 0.0560$.

Specific instructions follow this five-step procedure.

b. Determine whether or not the p -value is smaller than α .

The p -value is not smaller than the level of significance, α (0.05).

(ii) Using the classical procedure:

a. Determine the critical region and critical value(s).

The critical region is split into two equal parts because H_a expresses concern for values related to “different from”. The critical values are obtained from Statistical Table 6 at the intersections of row $df = 11$ with columns $\alpha = 0.975$ and 0.025 (area to right): $\chi^2(11, 0.975) = 3.82$ and $\chi^2(11, 0.025) = 21.9$.

For specific instructions, see Section 9.3 ahead.

b. Determine whether or not the calculated test statistic is in the critical region.

χ^2^* is not in the critical region; see Figure 9.31.

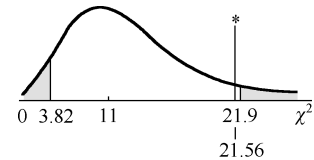


Figure 9.31 Finding p -value using the classical procedure

Step 5 The Results:

a. State the decision about H_o : Fail to reject H_o .

b. State the conclusion about H_a

There is not sufficient evidence at the 0.05 significance level to conclude that the shelf life of this hemical has a standard deviation different from 10 days. Therefore, Fast Photo should purchase the chemical accordingly.

Calculating the p -value when using the χ^2 -distribution

Method 1: Use Statistical Table 6 in Appendix to place bounds on the p -value. Inspect the $df=11$ row to locate $\chi^2^* = 21.56$. Notice that 21.56 is between two table entries, see Table 9.8. The bounds for $\frac{1}{2}P$ are read from the Right-Hand Tail heading at the top of the table.

Table 9.8 Portion of Statistical Table 6

df	Area in Right-Hand Tail			
	...	0.05	$\frac{1}{2}P$	
...				
11		19.7	21.56	21.9

$$0.025 < \frac{1}{2}P < 0.05$$

Double both bounds to find the bounds for P : $2 \times (0.025 < \frac{1}{2}P < 0.05)$ becomes $0.05 < P < 0.10$.

Note: When sample data are skewed, just one outlier can greatly affect the standard deviation, it is very important, especially when using small samples, that the sampled population be normal; otherwise, the procedures are not reliable.

New Words and Expressions

overfill ['əʊvə'fɪl] *vi.* 把.....装得溢出

underfill ['ʌndə'fɪl] *vi.* 未装满, 未充满

scenario [sə'nɑ:riəʊ] *n.* (行动的) 方案; 剧情概要; 分镜头剧本

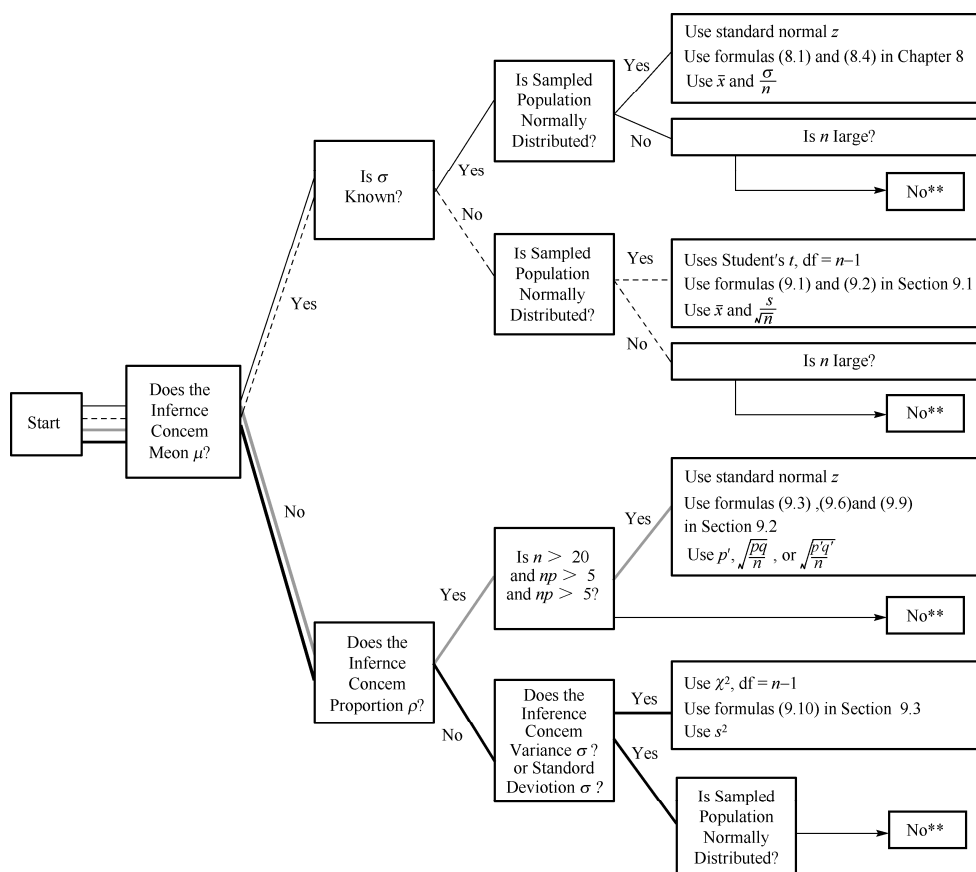
chemical ['kemɪkl] *n.* 化学药品, 化学制品

shelf life [ʃelf laɪf] *n.* (包装食品的) 货架期, 保存限期

promotion [prə'məʊʃn] *n.* 促进, 增进; 提升, 升级; (商品等的) 推广

Summary

We have been studying inferences, both confidence intervals and hypothesis tests, for the three basic population parameters (mean μ , proportion p , and standard deviation σ) of a single population. Most inferences about a single population are concerned with one of these three parameters. The following Figure 9.32 presents a visual organization of the techniques presented in Units 8 and 9 along with the key questions that you must ask as you are deciding which test statistic and formula to use.



No** means that a nonparametric technique (normal distribution not required) is used.

Figure 9.32 The techniques presented in Units 8 and 9.

In this unit we also used the maximum error of estimate, formula (9.7), to determine the size of the sample required to make estimations about the population proportion with the desired accuracy. By combining the reported point estimate and the sample size, we can determine the corresponding binomial proportion maximum error of estimate.

In the next unit we will discuss inferences about two populations whose respective means, proportions, and standard deviations are to be compared.

Reading English Materials

The History of Statistics: Student's t -statistic.

To many of us, whether statistician or not, the name William Sealy Gosset may be unrecognizable. His pseudonym Student, however, reveals him as one of the most prominent statisticians in history. Student's t -test is an important part of every introductory statistics course, making everyone from single-statistics-course students to those who have devoted their lives to the discipline familiar with his work.

Gosset was born in Canterbury in 1876, and studied chemistry and mathematics at New College, Oxford. After university, William was hired by Arthur Guinness, Son & Co. as a brewer at the St. James' Gate brewery in Dublin, where he worked from 1899 until 1935. At the time, Guinness became interested in hiring scientists who could apply their skills to the brewing process, and Gosset did not disappoint. In 1904, he wrote an internal report titled *The Application of the Law of Error to the work of the Brewery* where he made a case for introducing statistical methodologies to the brewing industry (Pearson, 1939). In fact, Gosset's first paper was an application of the Poisson distribution to yeast counts (Student, 1907).

In the conclusion of Gosset's report, he suggested consulting a "mathematical physicist" to address some of the more theoretical concerns. In 1906, he took a leave of absence from the brewery to study in the Biometric Lab of Karl Pearson. During this time, Gosset learned about distributional theory and the correlation coefficient. However, the large-sample theory that was made available to Gosset was not entirely practical to his work at the brewery; he seldom had the appropriately "large" sample sizes available to satisfy the assumptions of these methods.

This lack of small-sample methodology led to Gosset's most famous work in which he summarized the first four moments of the sample variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n}$$

and noticed their striking similarity to a Pearson Type III curve (Student, 1908a). His paper contains the derivation of the t -test (though not in its current form), some empirical work, examples and a statistical table for general use. However, the t -test would not see much use outside of his own brewery for many years.

Eventually, a young statistician named Ronald Fisher wrote to Gosset about the denominator of his sample variance; why was it not $(n-1)$? When Gosset asked Pearson about this, Pearson replied that n or $(n-1)$ means little in large samples but only to "naughty brewers" that "take n so small that the difference is not of the order of probable error" (Pearson, 1939). This initial exchange led to a lifelong friendship between Gosset and Fisher. Fisher thought highly of Gosset's work and eventually reparameterized Gosset's derivation into the familiar t distribution with corresponding degrees of freedom we know today. It is also perhaps, through Fisher's insistence and promotion of the work, that the method found itself in more general use outside of the brewery.

Gosset wrote a companion manuscript to his 1908 paper for the correlation coefficient (Student, 1908b), made contributions to the design and analysis of agricultural experiments, and later published papers in support of the theory of natural selection. In 1935, he moved to London to be head brewer at the new Guinness Park Royal brewery. He died in 1937 at the age of 61, survived by his wife, three children and one grandson. He published 22 manuscripts.



William Sealy Gosset, 1899

How did William Sealy Gosset become known as Student?

Perhaps out of fear of losing a competitive advantage, the brewery enforced a rule that forbade its scientists from publishing their research. Gosset argued that his work would be of no benefit to other brewers and was finally allowed to publish using a pseudonym – Student – to prevent other employees from noticing. It is interesting to note that two other chemists from the brewery published statistical work under assumed names: Sophister and Mathetes (Hotelling, 1930).

From the manuscripts listed below, it is possible to develop a very accurate picture of William Sealy Gosset. He was well-liked and respected by several notable statisticians including R.A. Fisher, Karl and Egon Pearson (Fisher, 1939; Pearson, 1939). He was a modest man, downplaying the importance of his work to the point where he declared “Fisher would have discovered it all anyway” (Boland, 1984). There is an interesting account of Gosset and Fisher’s relationship (not always free from statistical argument) described through the latter’s second eldest daughter (Box, 1981). McMullen, a former brewery coworker who marveled at Gosset’s many accomplishments, wrote a touching piece that describes Gosset’s personality and many interests in gardening, boat-building, biking, golfing, sailing and fishing (1939). Many of the aforementioned articles contain excerpts from Gosset’s letters to and from Fisher and Karl Pearson and illustrate his good sense of humor (Boland, 1984; Box, 1981; Pearson, 1939).

References

1. Boland P. J. (1984). A biographical glimpse of William Sealy Gosset. *The American Statistician* 38: 179-183.
2. Box J. F. (1981). Gosset, Fisher, and the t distribution. *The American Statistician* 35: 61-66
3. Hotelling H. (1930). British statistics and statisticians today. *Journal of the American Statistical Association* 25: 186-190.
4. Fisher R. A. (1939). Student. *Annals of Eugenics* 9: 1-9.
5. McMullen L. (1939). “Student” as a man. *Biometrika* 30: 205-210.
6. Pearson E. S. (1939). “Student” as a statistician. *Biometrika* 30: 210-250.
7. Student. (1907). On the error of counting with a haemocytometer. *Biometrika* 5: 351-360.
8. Student. (1908a). The probable error of a mean. *Biometrika* 6: 1-25.
9. Student. (1908b). The probable error of a correlation coefficient. *Biometrika* 6: 302-310.

New Words and Expressions

Arthur Guinness, Son & Co 阿瑟·健力士公司, 健力士啤酒 (Guinness) 是全球著名的黑啤, 是由阿瑟·吉尼斯 (Arthur Guinness) 于 1759 年在爱尔兰都柏林建立的一家酿酒厂。

brewer ['bru:ə(r)] *n.* 啤酒制造者

manuscript ['mænɪskript] *n.* 手稿; 原稿; (印刷术发明以前书籍或文献的) 手写本

lifelong ['laɪflɒŋ] *adj.* 毕生的, 终身的; 永生不渝的

forbade [fə'baed] *v.* 禁止 (forbid 的过去式); 妨碍; 阻碍; 阻止

employee [ɪm'plɔɪi:] *n.* 雇工, 雇员, 职工

coworker ['kəʊ,wɜ:kə] *n.* 共同工作的人, 同事, 合作者

marvel ['mɑ:vəl] *vt. & vi.* 惊奇, 对……感到惊奇

boat-building 游艇制造, 造船

bike [baɪk] *vi.* 骑自行车 *n.* 自行车; 摩托车; 电动自行车

aforementioned [ə,fɔ:'menʃənd] *adj.* 上述的; 前述的

Problems

9.1 Using Statistical Table 4 in Appendix, Find:

- a. $t(12, 0.01)$ b. $t(22, 0.025)$ c. $t(50, 0.10)$ d. $t(8, 0.005)$

9.2 A survey of 3000 randomly selected Minnesotans aged 65 and older revealed that, on average, they spent \$85 per month on prescription drugs, with a standard deviation of \$50.35 per month. Construct a 99% confidence interval for the true mean amount spent per month.

9.3 While writing an article on the high cost of college education, a reporter took a random sample of the cost of new textbooks for a semester. The random variable x is the cost of one book. Her sample data can be summarized by $n = 41$, $\bar{x} = 3582.17$, and $\sum (x - \bar{x})^2 = 9960.336$.

- a. Find the sample mean, \bar{x} . b. Find the sample standard deviation, s .
c. Find the 90% confidence interval to estimate the true mean textbook cost for the semester based on this sample.

9.4 The fuel economy information on a new SUV window sticker indicates that its new owner can expect 16 mpg (miles per gallon) in city driving and 20 mpg for highway driving and 18 mpg overall. Accurate gasoline records for one such vehicle were kept, and a random sample of mileage per tank of gasoline was collected:

17.6	17.7	18.1	22.0	17.0	19.4	18.9	17.4	21.0	19.2
18.3	19.1	20.7	16.7	19.4	18.2	18.4	17.1	17.4	15.8
17.9	18.0	16.3	17.5	17.3	20.4	19.1	21.0	18.1	19.0
19.6	18.9	16.8	18.2	17.6	19.1	18.0	16.8	20.9	17.9
17.7	20.3	18.6	19.0	16.5	19.4	18.6	18.6	17.3	18.7

- a. Determine whether an assumption of normality is reasonable, explain.
- b. Construct a 95% confidence interval for the estimate of the mean mileage per gallon.
- c. What does the confidence interval suggest about SUVs' fuel economy expectations as expressed on the window sticker?

9.5 State the null hypothesis, H_o , and the alternative hypothesis, H_a , that would be used to test each of the following claims:

- a. A chicken farmer at Best Broilers claims that his chickens have a mean weight of 56 oz.
- b. The mean age of U.S. commercial jets is less than 18 years.
- c. The mean monthly unpaid balance on credit card accounts is more than \$400.

9.6 Determine the p -value for the following hypothesis tests involving Student's t -distribution with 10 degrees of freedom.

- a. $H_o: \mu = 15.5, H_a: \mu \neq 15.5, t^* = -2.01$
- b. $H_o: \mu = 15.5, H_a: \mu \neq 15.5, t^* = 2.01$
- c. $H_o: \mu = 15.5, H_a: \mu \neq 15.5, t^* = 2.01$
- d. $H_o: \mu = 15.5, H_a: \mu \neq 15.5, t^* = -2.01$

9.7 Homes in a nearby college town have a mean value of \$88,950. It is assumed that homes in the vicinity of the college have a higher mean value. To test this theory, a random sample of 12 homes is chosen from the college area. Their mean valuation is \$92,460, and the standard deviation is \$5,200. Complete a hypothesis test using $\alpha = 0.05$. Assume prices are normally distributed.

- a. Solve using the p -value approach.
- b. Solve using the classical approach.

9.8 A bank randomly selected 250 checking account customers and found that 110 of them also had savings accounts at this same bank. Construct a 95% confidence interval for the true proportion of checking account customers who also have savings accounts.

9.9 Calculate the test statistic z , used in testing the following:

- a. $H_o: p = 0.70$ vs. $H_a: p > 0.70$, with the sample $n = 300$ and $x = 224$
- b. $H_o: p = 0.50$ vs. $H_a: p < 0.50$, with the sample $n = 450$ and $x = 207$
- c. $H_o: p = 0.35$ vs. $H_a: p = 0.35$, with the sample $n = 280$ and $x = 94$
- d. $H_o: p = 0.90$ vs. $H_a: p > 0.90$, with the sample $n = 550$ and $x = 508$

9.10 An insurance company states that 90% of its claims are settled within 30 days. A consumer group selected a random sample of 75 of the company's claims to test this statement. If the consumer group found that 55 of the claims were settled within 30 days, do they have sufficient reason to support their contention that less than 90% of the claims are settled within 30 days? Use $\alpha = 0.05$.

- a. Solve using the p -value approach.
- b. Solve using the classical approach.

9.11 a. If x successes result from a binomial experiment with $n = 1000$ and $p = P$ (success), and the 95% confidence interval for the true probability of success is determined, what is the maximum value possible for the "maximum error of estimate"?

- b. Explain how the results of national polls, like those from reputable polling establishments

such as Harris and Gallup, are related (similarities and differences) to the confidence interval technique studied in this section.

c. The theoretical sampling error with a level of confidence can be calculated, but polls typically report only a “margin of error” with no probability (level of confidence). Why is that?

9.12 Has the law requiring bike helmet use failed? Yankelovich Partners conducted a survey of bicycle riders in the United States. Only 60% of the nationally representative sample of 1020 bike riders reported owning a bike helmet.

a. Find the 95% confidence interval for the true proportion p for a binomial experiment of 1020 trials that resulted in an observed proportion of 0.60. Use this to estimate the percentage of bike riders who report owning a helmet.

b. Based on the survey results, would you say there is compliance with the law requiring bike helmet use? Explain.

Suppose you wish to conduct a survey in your city to determine what percent of bicyclists own helmets. Use the national figure of 60% for your initial estimate of p .

c. Find the sample size if you want your estimate to be within 0.02 with 95% confidence.

d. Find the sample size if you want your estimate to be within 0.04 with 95% confidence.

e. Find the sample size if you want your estimate to be within 0.02 with 90% confidence.

f. What effect does changing the maximum error have on the sample size? Explain.

g. What effect does changing the level of confidence have on the sample size? Explain.

9.13 Find these critical values by using Table 8 of Appendix B.

a. $\chi^2(18, 0.01)$

b. $\chi^2(16, 0.025)$

c. $\chi^2(8, 0.10)$

d. $\chi^2(28, 0.01)$

e. $\chi^2(22, 0.95)$

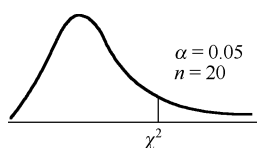
f. $\chi^2(10, 0.975)$

g. $\chi^2(50, 0.90)$

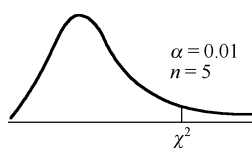
h. $\chi^2(24, 0.99)$

9.14 Using the notation of problem 9.33, name and find the critical values of χ^2 .

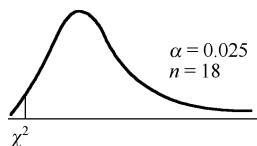
a.



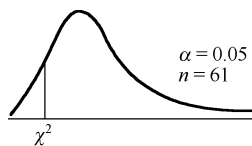
b.



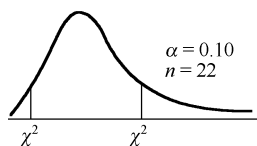
c.



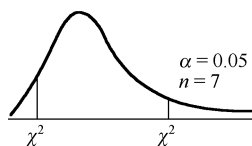
d.



e.



f.



9.15 Determine the critical region and critical value(s) that would be used to test the following using the classical approach:

- a. $H_0: \sigma = 0.5$ and $H_a: \sigma > 0.5$, with $n = 18$ and $\alpha = 0.05$
- b. $H_0: \sigma^2 = 8.5$ and $H_a: \sigma^2 < 8.5$, with $n = 15$ and $\alpha = 0.01$
- c. $H_0: \sigma = 20.3$ and $H_a: \sigma \neq 20.3$, with $n = 10$ and $\alpha = 0.10$
- d. $H_0: \sigma^2 = 0.05$ and $H_a: \sigma^2 \neq 0.05$, with $n = 8$ and $\alpha = 0.02$
- e. $H_0: \sigma = 0.5$ and $H_a: \sigma < 0.5$, with $n = 12$ and $\alpha = 0.10$

9.16 A random sample of 51 observations was selected from a normally distributed population. The sample mean was $\bar{x} = 98.2$, and the sample variance was $s^2 = 37.5$. Does this sample show sufficient reason to conclude that the population standard deviation is not equal to 8 at the 0.05 level of significance?

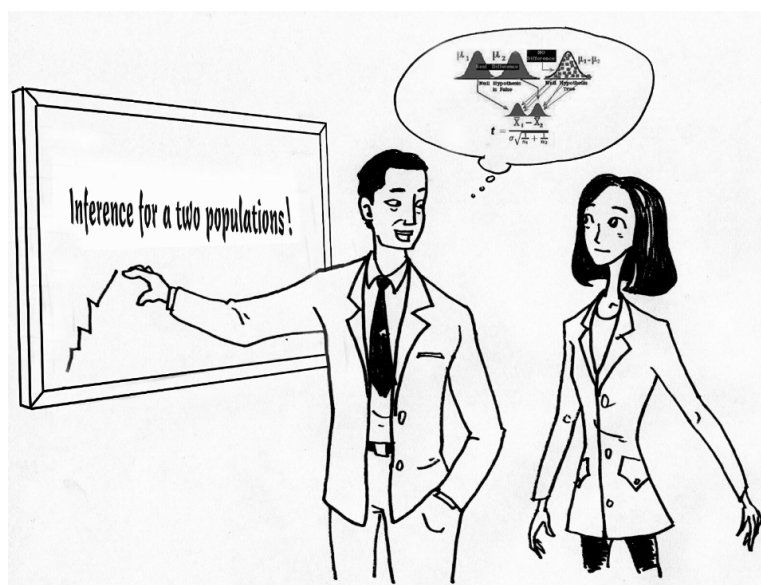
- a. Solve using the p -value approach.
- b. Solve using the classical approach.

9.17 In the past the standard deviation of weights of certain 32.0-oz packages filled by a machine was 0.25 oz. A random sample of 20 packages showed a standard deviation of 0.35 oz. Is the apparent increase in variability significant at the 0.10 level of significance? Assume package weight is normally distributed.

- a. Solve using the p -value approach.
- b. Solve using the classical approach.

The best thing about being a statistician is that you get to play in everybody else's backyard.

—John Tukey



Unit 10

Inferences Involving Two Populations



10.1 Dependent and Independent Samples



10.2 Inferences Concerning the Mean Difference Using Two
Dependent Samples



10.3 Inferences Concerning the Difference between Means Using
Two Independent Samples



10.4 Inferences Concerning the Difference between Proportions
Using Two Independent Samples



10.5 Inferences Concerning the Ratio of Variances Using Two
Independent Samples



Summary



Problems

10.1 Dependent and Independent Samples

In this unit we are doing to study the procedures for making inferences about populations.

Figure 10.1 gives you an overview of how the procedures will unfold over the course of the unit.

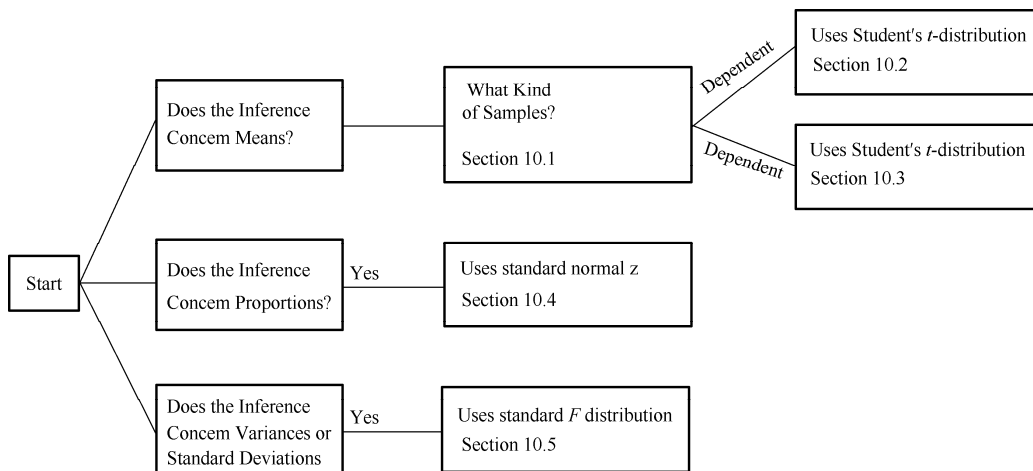


Figure 10.1 “Road Map” to two population inferences

When comparing two populations, we need two samples, one from each population. Two basic kinds of samples can be used: independent and dependent. The dependence or independence of two samples is determined by the sources of the data. A **data source** can be a person, an object, or anything that yields a piece of data. If the same set of sources or related sets are used to obtain the data representing both populations, we have **dependent samples**. If two unrelated sets of sources are used, one set from each population, we have **independent samples**.

Example 10.1

To get a better sense of the differences between dependent and independent samples, imagine that you are conducting a test to see whether the participants in a physical fitness class actually improve in their level of fitness. It is anticipated that approximately 500 people will sign up for this course. You decide to give 50 of the participants a set of tests before the course begins (a pretest), and then give another set of tests to 50 participants at the end of the course (a posttest). Two sampling procedures are proposed:

Plan A: Randomly select 50 participants from the list of those enrolled and give them the pretest. At the end of the course, make a second random selection of size 50 and give them the posttest.

Plan B: Randomly select 50 participants and give them the pretest; give the same set of 50 the posttest when they complete the course.

Plan A illustrates independent sampling; the sources (the class participants) used for each sample (pretest and posttest) were selected separately. Plan B illustrates dependent sampling; the sources used for both samples (pretest and posttest) are the same.

Typically, when both a pretest and a posttest are used, the same subjects participate in the study. Thus, pretest versus posttest (before versus after) studies usually use dependent samples. Studies can, however, also use before versus after examinations in conjunction with independent samples.

Example 10.2

We consider this test that is being designed to compare the wearing quality of two brands of automobile tires. The automobiles will be selected and equipped with the new tires and then driven under “normal” conditions for one month. Then a measurement will be taken to determine how much wear took place. Two plans are proposed:

Plan C: A sample of cars will be selected randomly, equipped with brand A tires, and driven for the month. Another sample of cars will be selected, equipped with brand B tires, and driven for the month.

Plan D: A sample of cars will be selected randomly, equipped with one tire of brand A and one tire of brand B (the other two tires are not part of the test), and driven for the month.

We suspect that many other factors must be taken into account when testing automobile tires—such as age, weight, and mechanical condition of the car; driving habits of drivers; location of the tire on the car; and where and how much the car is driven. However, at this time we are trying only to illustrate dependent and independent samples. Plan C is independent (unrelated sources), and plan D is dependent (common sources).

Independent and dependent samples each have their advantages; these will be emphasized later. Both methods of sampling are often used.

Definition 1

■ **(Data) Source:** A person, an object, or anything that yields a piece of data.

Definition 2

■ **Dependent samples:** If the same set of sources or related sets are used to obtain the data representing both populations, we have dependent samples.

■ **Independent samples:** If two unrelated setsof sources are used, one set from each population, we have independent samples.

New Words and Expressions

unfold [ʌn'fəʊld] *vt.* 摊开；展现，披露 *vi.* 逐渐显露；开展，发展

sign [saɪn] *vt.* 和……签约（或应聘）；示意；记下，记录 sign up 签订，报名参加

pretest ['pri:tɛst] *n.* 预备考试，预备调查

posttest ['pəʊstɪst] *n.* 课程结束考核

Technical Terms

data source 数据来源，数据源

dependent samples 相关样本
independent samples 独立样本

Notes

1. 同义词辨析: unfold, open 这两个动词均有“打开”之意。
unfold: 主要指把原来包好、卷好或叠好的东西再打开。
open: 普通用词, 指把原来关起来或盖紧的东西打开。

10.2 Inferences Concerning the Mean Difference Using Two Dependent Samples

The procedures for comparing two population means are based on the relationship between two sets of sample data, both samples from the same or related sources.

When dependent samples are involved, the data are thought of as “paired data”. The data may be paired as a result of being obtained from “before” and “after” studies; from pairs of identical twins; from a “common” source, as with the amounts of tire wear for each brand in plan D in Section 10.1; or from matching two subjects with similar traits to form “matched pairs.” The pairs of data values are compared directly to each other by using the difference in their numerical values. The resulting difference is called a *paired difference*.

Paired Difference

$$d = x_1 - x_2 \quad (10.1)$$

Example 10.3 (Example 10.2 continued)

Using paired data this way has a built-in ability to remove the effect of otherwise uncontrolled factors. The tire-wear problem using plan C and plan D is an excellent example of such additional factors. The wearing ability of a tire is greatly affected by a multitude of factors: the size, weight, age, and condition of the car, the driving habits of the driver, the number of miles driven, the condition and types of roads driven on, the quality of the material used to make the tire, and so on. With plan D, we create paired data by mounting one tire from each brand on the same car. Since one tire of each brand will be tested under the same conditions, same car, same driver, and so on, the extraneous causes of wear are neutralized.

10.2.1 Procedures and Assumptions for Inferences Involving Paired Data

Example 10.4 (Example 10.2 continued)

The test comparing the wear of tires from two different tire companies uses plan D as described in section 10.1. All the aforementioned factors will have an equal effect on both brands

of tires, car by car. The test places one tire of each brand on each of the six test cars. The position (left or right side, front or back) was determined with the aid of a random-number table. Table 10.1 lists the resulting amounts of wear (in thousandths of an inch).

Table 10.1 Amount of Tire Wear

Car	1	2	3	4	5	6
Brand A	125	64	94	38	90	106
Brand B	133	65	103	37	102	115

Since the various cars, drivers, and conditions are the same for each tire of a paired set of data, it makes sense to use a third variable, the paired difference d . Our two dependent samples of data may be combined into one set of d values, where $d = B - A$.

Table 10.2 Value of Paired Difference

Car	1	2	3	4	5	6
$d = B - A$	8	1	9	-1	12	9

The difference between the two population means, when dependent samples are used (often called “**dependent means**”), is equivalent to the **mean of the paired differences**. Therefore, when an inference is to be made about the difference of two means and paired differences are used, the inference will in fact be about the mean of the paired differences. The sample mean of the paired differences will be used as the point estimate for these inferences.

In order to make inferences about the mean μ_d of all possible paired differences, we need to know about the *sampling distribution* of \bar{d} .

When paired observations are randomly selected from normal populations:

The paired difference, $d = x_1 - x_2$, will be approximately normally distributed about a mean μ_d with a standard deviation of σ_d .

This is another situation in which the t -test for one mean is applied; namely, we wish to make inferences about an unknown mean (μ_d) where the random variable (d) involved has an approximately normal distribution with an unknown standard deviation (σ_d).

Inferences about the mean of all possible paired differences μ_d are based on samples of n dependent pairs of data and the t -distribution with $n-1$ degrees of freedom, under the following assumption:

Assumption for inferences about the mean of paired differences μ_d :

The paired data are randomly selected from normally distributed populations.

10.2.2 Confidence Interval Procedure

The $1-\alpha$ confidence interval for estimating the mean difference μ_d is found using this formula:

Confidence Interval for μ_d

$$\bar{d} - t(df, \alpha/2) \cdot \frac{s_d}{\sqrt{n}} \text{ to } \bar{d} + t(df, \alpha/2) \cdot \frac{s_d}{\sqrt{n}}, \quad \text{where } df = n - 1 \quad (10.2)$$

\bar{d} is the mean of the sample differences:

$$\bar{d} = \frac{\sum d}{n} \quad (10.3)$$

and s_d is the standard deviation of the sample differences:

$$\bar{d} = \frac{\sqrt{\sum d^2 - \left[\frac{(\sum d)^2}{n} \right]}}{n-1} \quad (10.4)$$

Constructing a Confidence Interval for μ_d

Example 10.5 (Example 10.2 continued)

To illustrate how the confidence interval for μ_d can be formed, we will use the paired data on tire wear as reported in Table 10.1 and assume the amounts of wear are approximately normally distributed for both brands of tires. Using the five-step process, we can construct the 95% confidence interval for the mean difference in the paired data. The sample information is $n = 6$ pieces of paired data, $\bar{d} = 6.3$, and $s_d = 5.1$.

Step 1 Parameter of interest:

μ_d , the mean difference in the amounts of wear between the two brands of tires.

Step 2 a. Assumptions:

Both sampled populations are approximately normal.

b. Probability distribution:

The t -distribution with $df = 6 - 1 = 5$ and formula (10.2) will be used.

c. State the level of confidence:

$$1 - \alpha = 0.95.$$

Step 3 Sample information:

$$n = 6, \bar{d} = 6.3, \text{ and } s_d = 5.1.$$

Step 4 a. Confidence coefficient:

This is a two-tailed situation with $\alpha/2 = 0.025$ in one tail. From Table 4 in Appendix Tables, $t(df, \alpha/2) = t(5, 0.025) = 2.57$.

b. Maximum error of estimate:

Using the maximum error part of formula (10.2), we have

$$E = t(df, \alpha/2) \cdot \frac{s_d}{\sqrt{n}} :$$

$$E = 2.57 \cdot \left(\frac{5.1}{\sqrt{6}} \right) = (2.57)(2.082) = 5.351 = 5.4$$

c. Lower/upper confidence limits:

$$\bar{d} \pm E$$

$$6.3 \pm 5.4$$

$$6.3 - 5.4 = 0.9 \text{ to } 6.3 + 5.4 = 11.7$$

Step 5 a. Confidence interval: 0.9 to 11.7 is the 95% confidence interval for μ_d .

b. That is, with 95% confidence we can say that the mean difference in the amounts of wear is between 0.9 and 11.7 thousandths of an inch. Or, in other words, the population mean tire wear from Brand B is between 0.9 and 11.7 thousandth of an inch greater than the population mean tire wear for Brand A.

This is quite a wide confidence interval, in part because of the small sample size. Recall from the central limit theorem that as the sample size increases, the standard error (estimated by s_d / \sqrt{n}) decreases.

10.2.3 Hypothesis-Testing Procedure

When we test a null hypothesis about the mean difference, the test statistic used will be the difference between the sample mean \bar{d} and the hypothesized value of μ_d , divided by the estimated standard error. This statistic is assumed to have a t -distribution when the null hypothesis is true and the assumptions for the test are satisfied. The value of the test statistic t^* is calculated as follows:

Test Statistic for μ_d

$$t^* = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}, \quad \text{where } df = n - 1 \quad (10.5)$$

Note: A hypothesized mean difference, μ_d , can be any specified value. The most common value specified is zero; however, the difference can be nonzero.

One-Tailed Hypothesis Test for μ_d

Example 10.6

Let's conduct a one-tailed hypothesis test for the mean difference by looking at a study on high blood pressure and the drugs used to control it. The effect of calcium channel blockers on pulse rate was one of many specific concerns in the study. Twenty-six patients were randomly selected from a large pool of potential subjects, and their pulse rates were recorded. A calcium channel blocker was administered to each patient for a fixed period of time, and then each patient's pulse rate was again determined. The two resulting sets of data appeared to have approximately normal distributions, and the statistics were $\bar{d} = 1.07$ and $s_d = 1.74$ (d = before–after). Using the five-step method, can we determine if the sample information provides sufficient evidence to show that this calcium channel blocker lowered the pulse rate? In other words, if “lower pulse rate” means that “after” is less than “before, then before–after” should be positive. Does the sample information provide confirming evidence? Use $\alpha = 0.05$.

Step 1

a. Parameter of interest:

μ_d , the mean difference (reduction) in pulse rate from before to after using the calcium channel blocker for the time period of the test.

b. Statement of hypotheses:

$$H_o: \mu_d = 0 (\leq) \text{ (did not lower rate)}$$

Remember: d = before–after.

$$H_a: \mu_d > 0 \text{ (did lower rate)}$$

Step 2

a. Assumptions: Since the data in both sets are approximately normal, it seems reasonable to assume that the two populations are approximately normally distributed.

b. Test statistic: The t -distribution with $df = n-1 = 25$, and the test statistic is t^* from formula (10.5).

c. Level of significance: $\alpha = 0.05$.

Step 3 a. Sample information: $n = 26$, $\bar{d} = 1.07$, and $s_d = 1.74$.

b. Calculate the value of the test statistic.

$$t^* = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}} : t^* = \frac{1.07 - 0.0}{1.74 / \sqrt{26}} = \frac{1.07}{0.34} = 3.14$$

Step 4 The probability distribution:

As always, we can use either the p -value or the classical procedure:

(i) Using the p -value procedure:

a. Use the right-hand tail because H_a expresses concern for values related to “greater than.” $P = P(t^* > 3.14, \text{ with } df = 25)$ as shown in Figure 10.2.

To find the p -value, you have three options:

1. Use Statistical Table 4 in Appendix to place bounds on the p -value: $P < 0.005$.
2. Use Statistical Table 5 in Appendix to read the value directly: $P = 0.002$.
3. Use a computer or calculator to find the p -value: $P = 0.0022$.

Specific instructions are in Section 9.1.

b. The p -value is smaller than the level of significance, α .

(ii) Using the classical procedure:

a. The critical region is the right-hand tail because H_a expresses concern for values related to “greater than”. The critical value is obtained from Statistical Table 4: $t(25, 0.05) = 1.71$, see Figure 10.3.

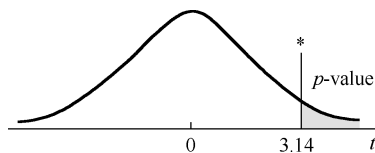


Figure 10.2 Finding p -value

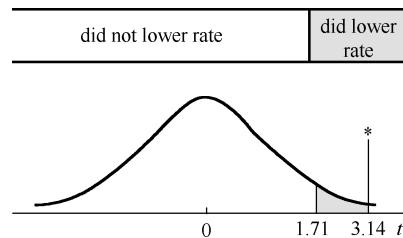


Figure 10.3 From Statistical Table 4

Specific instructions are in Section 9.1.

b. t^* is in the critical region, as shown in black on Figure 10.3.

Step 5

a. Decision: Reject H_o .

b. Conclusion: At the 0.05 level of significance, we can conclude that the average pulse rate is lower after the administration of the calcium channel blocker.

In the preceding detailed hypothesis test, the results showed a statistical significance with a p -value of 0.002—that is, 2 chances in 1,000. We can see that “*statistical significance*” does not always have the same meaning when the “practical” application of the results is considered. A more practical question might be: Is lowering the pulse rate by this small average amount, estimated to be 1.07 beats per minute, worth the risks of possible side effects of this medication? Actually, the whole issue is much broader than just this one issue of pulse rate.

Two-Tailed Hypothesis Test for μ_d

Example 10.7 (Example 10.2 continued)

To conduct this two-tailed hypothesis test, let’s return to the data we collected from two different brands of tires. Suppose the sample data in Table 10.1 were collected with the hope of showing that the two brands do not wear equally. Assuming the amounts of wear are approximately normally distributed, we will use the five-step process to determine whether the data provide sufficient evidence to conclude that the two brands show unequal wear at the 0.05 level of significance.

Step 1

a. Parameter of interest: μ_d , the mean difference in the amounts of wear between the two brands.

b. State the null hypothesis (H_o) and the alternative hypothesis (H_a):

$$H_o: \mu_d = 0 \text{ (no difference)}$$

Remember: $d = B - A$.

$$H_a: \mu_d \neq 0 \text{ (difference)}$$

Step 2 a. Assumptions: e assumption o normality is included in the statement of this problem.

b. Test statistic: The t -distribution with

$$df = n - 1 = 6 - 1 = 5, \text{ and } t^* = (\bar{d} - \mu_d) / (s_d / \sqrt{n}).$$

c. Level of significance: $\alpha = 0.05$.

Step 3 Sample information: $n = 6$, $\bar{d} = 6.3$, and $s_d = 5.1$.

b. Calculated test statistic:

$$t^* = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}};$$
$$t^* = \frac{6.3 - 0.0}{5.1 / \sqrt{6}} = \frac{6.3}{2.08} = 3.03$$

Step 4 The probability distribution:

Again, we can use either the p -value or the classical procedure:

(i) Using the *p*-value procedure

a. Use both tails because H_a expresses concern for values related to “different from”.

$$P = p\text{-value} = P(t^* < -3.03) + P(t^* > 3.03) = 2 \times P(|t^*| > 3.03)$$

as shown in Figure 10.4.

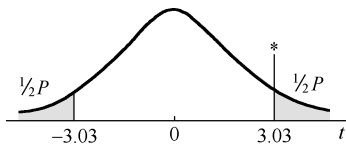


Figure 10.4 Finding *p*-value

To find the *p*-value, you have three options:

1. Use Statistical Table 4 in Appendix: $0.02 < P < 0.05$.

2. Use Statistical Table 5 in Appendix to place bounds on the

p-value: $0.026 < P < 0.030$.

3. Use a computer or calculator to find the *p*-value: $P = 2 \times 0.0145 = 0.0290$.

For specific instructions, see Section 9.1.

b. The *p*-value is smaller than α .

(ii) Using the classical procedure

a. The critical region is two-tailed because H_a expresses concern for values related to “different than”. The critical value is obtained from Statistical Table 4: $t(5, 0.025) = 2.57$, see Figure 10.5.

For specific instructions, see Section 9.1.

b. t^* is in the critical region, as shown in black on

Figure 10.5.

Step 5

a. Decision: Reject H_0

b. Conclusion: There is a significant difference in the mean amounts of wear at the 0.05 level of significance.

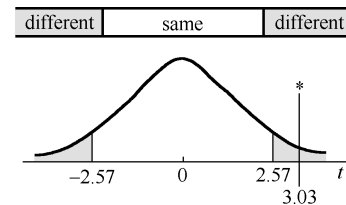


Figure 10.5 From Statistical Table 4

New Words and Expressions

remove [rɪ'mu:v] *vt.* 开除；去除；脱掉，拿下；迁移 *vi.* 迁移，移居；离开

calcium [kælsiəm] *n.* [化]钙

channel ['tʃænl] *n.* 频道，波道；渠道；途径；海峡

blockers [b'lɒkəz] *n.* 阻断剂；护航者；锥形锻模（blocker 的复数形式）

pulse rate [pʌls reɪt] 搏率，脉搏跳律

Technical Terms

paired difference 配对差

beats per minute 每分钟心跳次数

10.3 Inferences Concerning the Difference between Means Using Two Independent Samples

When comparing the means of two populations, we typically consider the difference between their means, $\mu_1 - \mu_2$ (often called “*independent means*”).

The inferences about $\mu_1 - \mu_2$ will be based on the difference between the observed sample means, $\bar{x}_1 - \bar{x}_2$. This observed difference, $\bar{x}_1 - \bar{x}_2$, belongs to a sampling distribution with the characteristics described in the following statement.

If independent samples of sizes n_1 and n_2 are drawn randomly from large populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$, the difference between the sample means, has

(i) mean $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$ and

(ii) standard error $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}$ (10.6)

If both populations have normal distributions, then the sampling distribution of $\bar{x}_1 - \bar{x}_2$ will also be normally distributed.

The “*t*-Distribution”

As head brewer at *Guinness Brewing Company*, William Gosset (aka Student) was faced with many small sets of data small by necessity because a 24-hour period often resulted in only one data value. Thus, he developed the *t*-test to handle these small samples for quality control in brewing. In his paper *The Probable Error of a Mean*, he set out to find the distribution of the amount of error in the sample mean, $(\bar{x} - \mu)$ divided by s , where s was from a sample of any known size. He then found the probable error of a mean, \bar{x} , for any size sample, by using the distribution of $(\bar{x} - \mu) / (s / \sqrt{n})$. Student’s *t*-distribution did not immediately gain popularity, and in fact in 1922, 14 years after the publication of his paper, Student wrote to noted statistician Ronald A. Fisher: “I am sending you a copy of Student’s Tables as you are the only man that’s ever likely to use them!” Today, Student’s *t*-distribution is widely used and respected in statistical research.

The preceding statement is true for all sample sizes, given that the populations involved are normal and the population variances σ_1^2 and σ_2^2 are known quantities. However, as with inferences about one mean, the variance of a population is generally an unknown quantity. Therefore, it will be necessary to estimate the standard error by replacing the variances, σ_1^2 and σ_2^2 , in formula (10.6) with the best estimates available—namely, the sample variances, s_1^2 and s_2^2 . The *estimated standard error* will be found using the following formula:

$$\text{estimated standard error} = \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \quad (10.7)$$

Inferences about the difference between two population means, $\mu_1 - \mu_2$ will be based on the following assumptions:

Assumption for inferences about the difference between two means, $\mu_1 - \mu_2$:

The samples are randomly selected from normally distributed populations, and the samples are selected in an independent manner.

No assumptions are made about the population variances.

Since the samples provide the information for determining the standard error, the ***t*-distribution** will be used as the test statistic. The inferences are divided into two cases.

Case 1: The *t*-distribution will be used, and the number of degrees of freedom will be calculated.

Case 2: The *t*-distribution will be used, and the number of degrees of freedom will be approximated.

Case 1 will occur when you are completing the inference using a computer or statistical calculator and the statistical software or program calculates the number of degrees of freedom for you. The calculated value for df is a function of both sample sizes and their relative sizes, and both sample variances and their relative sizes. The value of df will be a number between the smaller of $df_1 = n_1 - 1$ or $df_2 = n_2 - 1$ and the sum of the degrees of freedom, $df_1 + df_2 = [(n_1 - 1) + (n_2 - 1)] = n_1 + n_2 - 2$.

Case 2 will occur when you are completing the inference *without the aid of a computer or calculator and its statistical software package*. Use of the *t*-distribution with the smaller of $df_1 = n_1 - 1$ or $df_2 = n_2 - 1$ will give conservative results. Because of this approximation, the true level of confidence for an interval estimate will be slightly higher than the reported level of confidence, or the true *p*-value and the true level of significance for a hypothesis test will be slightly less than reported. The gap between these reported values and the true values will be quite small, unless the sample sizes are quite small and unequal or the sample variances are very different. The gap will decrease as the samples increase in size or as the sample variances are more alike.

Note: $A > B$ (“*A* is greater than *B*”) is equivalent to $B < A$ (“*B* is less than *A*”). When the difference between *A* and *B* is being discussed, it is customary to express the difference as “larger-smaller” so that the resulting difference is positive: $A - B > 0$. To express the difference as “smaller-larger” results in $B - A < 0$ (the difference is negative), which is usually unnecessarily confusing. Therefore, it is recommended that the difference be expressed as “larger-smaller”.

Since the only difference between the two cases is the number of degrees of freedom used to identify the *t*-distribution involved, we will study case 2 first.

10.3.1 Confidence Interval Procedure

We will use the following formula for calculating the endpoints of the $1 - \alpha$ confidence interval.

Confidence Interval for the Difference between Two Means (Independent Samples)

$$(\bar{x}_1 - \bar{x}_2) - t(df, \alpha/2) \cdot \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \quad \text{to} \quad (\bar{x}_1 - \bar{x}_2) + t(df, \alpha/2) \cdot \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)} \quad (10.8)$$

where df equals the smaller of df₁ or df₂

Example 10.8

In order to construct a confidence interval for the difference between two means, let's look at some sample information on student heights from a certain college campus. The sample information, given in Table 10.3, contains the heights (in inches) of 20 randomly selected women and 30 randomly selected men, taken in order to estimate the difference in their mean heights. We will assume that the heights are approximately normally distributed for both populations as we begin the five-step process to find the 95% confidence interval for the difference between the mean heights, $\mu_m - \mu_f$.

Table 10.3 Sample Information on Student Heights

Sample	Number	Mean	Standard Deviation
Female(<i>f</i>)	20	63.8	2.18
Male(<i>m</i>)	(<i>m</i>)	30	69.8 1.92

Step 1 Parameter of interest: $\mu_m - \mu_f$, the difference between the mean height of male students and the mean height of female students.

Step 2 a. Assumptions: Both populations are approximately normally distributed, and the samples were random and independently selected.

b. Probability distribution: The *t*-distribution with df = 19, the smaller of $n_m - 1 = 30 - 1 = 29$ or $n_f - 1 = 20 - 1 = 19$, and formula (10.8).

c. Level of confidence: $1 - \alpha = 0.95$.

Step 3 Sample information: See Table 10.2.

Step 4 a. Confidence coefficient: We have a two-tailed situation with $\alpha/2 = 0.025$ in one tail and df = 19. From Statistical Table 4 in Appendix, $t(df, \alpha/2) = t(19, 0.025) = 2.09$, see Figure 10.6.

See Section 9.1 for instructions on using Statistical Table 4.

b. Maximum error of estimate: Using the maximum error part of formula (10.8), we have

$$E = t(df, \alpha/2) \cdot \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

$$E = 2.09 \cdot \sqrt{\left(\frac{1.92^2}{30}\right) + \left(\frac{2.18^2}{20}\right)} = (2.09)(0.60) = 1.25$$

c. Lower and upper confidence limits.

$$(\bar{x}_1 - \bar{x}_2) \pm E$$

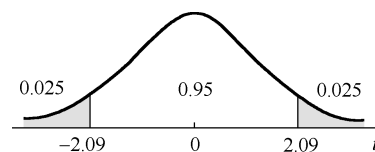


Figure 10.6 From Statistical Table 4

$$6.00 \pm 1.25$$

$$6.00 - 1.25 = 4.75 \quad \text{to} \quad 6.00 + 1.25 = 7.25$$

Step 5 a. Confidence interval: 4.75 to 7.25 is the 95% confidence interval for $\mu_m - \mu_f$.

b. That is, with 95% confidence, we can say that the difference between the mean heights of the male and female students is between 4.75 and 7.25 inches; that is, the mean height of male students is between 4.75 and 7.25 inches greater than the mean height of female students.

10.3.2 Hypothesis-Testing Procedure

When we test a null **hypothesis about the difference between two population means**, the test statistic used will be the difference between the observed difference of the sample means and the hypothesized difference of the population means, divided by the estimated standard error. The test statistic is assumed to have approximately a t -distribution when the null hypothesis is true and the normality assumption has been satisfied. The calculated value of the **test statistic** is found using this formula:

Test Statistic for the Difference between Two Means (Independent Samples)

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \quad (10.9)$$

where df is the smaller of df_1 or df_2

Note: A hypothesized difference between the two population means, $\mu_1 - \mu_2$, can be any specified value. The most common value specified is zero; however, the difference can be nonzero.

In order to examine the hypothesis procedure more closely, let's look at one- and two-tailed tests.

One-Tailed Hypothesis Test for the Difference Between Two Means

Example 10.9

For this test, let's suppose that we are interested in comparing the academic success of college students who belong to fraternal organizations with the academic success of those who do not belong to fraternal organizations. The reason for the comparison is the recent concern that fraternity members, on the average, are achieving at a lower academic level than nonfraternal students. (Cumulative grade-point average is used to measure academic success.) Random samples of size 40 are taken from each population and are listed in Table 10.4. Using the five-step process, let's complete a hypothesis test using $\alpha = 0.05$ and assume that the grade-point averages for both groups are approximately normally distributed.

Table 10.4 Sample Information on Academic Success

Sample	Number	Mean	Standard Deviation
Fraternity members (f)	40	2.03	0.68
Nonmembers (n)	40	2.21	0.59

Step 1 a. Parameter of interest: $\mu_n - \mu_f$ the difference between the mean grade-point averages for the nonfraternity members and the fraternity members.

b. Statement of hypotheses:

$H_o: \mu_n - \mu_f = 0 (\leq)$ (fraternity averages are no lower)

$H_a: \mu_n - \mu_f > 0$ (fraternity averages are lower)

Step 2 a. Assumptions: Both populations are approximately normally distributed, and random samples were selected. Since the two populations are separate, the samples are independent.

b. Test statistic: The t -distribution with df = the smaller of df_n , or df_f ; since both n 's are 40, $df = 40 - 1 = 39$; and t^* is calculated using formula (10.9).

c. Level of significance: $\alpha = 0.05$.

Step 3 a. Sample information: See Table 10.4.

b. Calculated test statistic (when df is not in the table, use the next smaller df value):

$$t^* = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}}$$

$$t^* = \frac{(2.21 - 2.03) - (0.00)}{\sqrt{\left(\frac{0.59^2}{40}\right) + \left(\frac{0.68^2}{40}\right)}} = \frac{0.18}{\sqrt{0.00870 + 0.1156}} = \frac{0.18}{0.1423} = 1.26$$

Step 4 The probability distribution:

Again, we can use either the p -value or the classical procedure:

(i) Using the p -value procedure:

a. Use the right-hand tail because H_a expresses concern for values related to “greater than”. $P = P(t^* > 1.26, \text{ with } df = 39)$ as shown in Figure 10.7.

To find the p -value, use one of three methods:

1. Use Statistical Table 4 in Appendix to place bounds on the p -value: $0.10 < P < 0.25$.
2. Use Statistical Table 5 in Appendix to place bounds on the p -value: $0.100 < P < 0.119$.
3. Use a computer or calculator to find the p -value: $P = 0.1076$.

Specific details follow this example.

b. The p -value is not smaller than α .

(ii) Using the classical procedure:

a. The critical region is the right-hand tail because H_a expresses concern for values related to “greater than”. The critical value is obtained from Statistical Table 4: $t(39, 0.05) = 1.69$, see Figure 10.8.

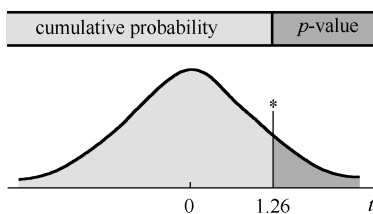


Figure 10.7 From Statistical Table 4 in Appendix

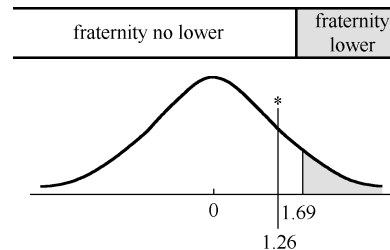


Figure 10.8 From Statistical Table 4 in Appendix

See Section 9.1 for information about critical values.

b. t^* is not in the critical region, as shown in black on Figure 10.8.

Step 5 a. Decision about H_o : Fail to reject H_o .

b. Conclusion: At the 0.05 level of significance, the claim that the fraternity members achieve at a lower level than nonmembers is not supported by the sample data.

To Find the p -Value Use One of Three Methods:

◆ Method 1: Use Statistical Table 4. Find 1.26 between two entries in the $df = 39$ row and read the bounds for P from the one-tail heading at the top of the table: $0.10 < P < 0.25$.

◆ Method 2, Use Statistical Table 5. Find $t^* = 1.26$ between two rows and $df = 39$ between two columns; read the bounds for $P(t^* > 1.26 | df = 39)$; $0.100 < P < 0.119$.

◆ Method 3: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value for you.

Two-Tailed Hypothesis for the Difference between Two Means

The difference between two means can also be used in a two-tailed situation.

Example 10.10

For example, many students have complained that the soft-drink vending machine A (in the student recreation room) dispenses a different amount of drink than machine B (in the faculty lounge). To test this belief, a student randomly sampled several servings from each machine and carefully measured them, with the results shown in Table 10.5.

Table 10.5 Sample Information on Vending Machines

Machine	Number	Mean	Standard Deviation
A	10	5.38	1.59
B	12	5.92	0.83

Does this evidence support the hypothesis that the mean amount dispensed by machine A is different from the amount dispensed by machine B? Let's assume that the amounts dispensed by both machines are normally distributed and complete the test using the five-step process, with $\alpha = 0.10$.

Step 1 a. Parameter of interest: $\mu_B - \mu_A$, the difference between the mean amount dispensed by machine B and the mean amount dispensed by machine A.

b. Statement of hypotheses:

H_o : $\mu_B - \mu_A = 0$ (A dispenses the same amount as B)

H_o : $\mu_B - \mu_A \neq 0$ (A dispenses a different amount than B)

Step 2 a. Assumptions: Both populations are assumed to be approximately normal, and the samples were random and independently selected.

b. Test statistic: The t -distribution with $df =$ the smaller of $n_A - 1 = 10 - 1 = 9$ or $n_B - 1 = 12 - 1 = 11$, so $df = 9$, and t^* calculated using formula (10.9).

c. Level of significance: $\alpha = 0.10$.

Step 3 a. Sample information: See Table 10.4.

b. Calculated test statistic:

$$t^* = \frac{(\bar{x}_B - \bar{x}_A) - (\mu_B - \mu_A)}{\sqrt{\left(\frac{s_B^2}{n_B}\right) + \left(\frac{s_A^2}{n_A}\right)}}$$
$$t^* = \frac{(5.92 - 5.38) - (0.00)}{\sqrt{\left(\frac{0.83^2}{12}\right) + \left(\frac{1.59^2}{10}\right)}} = \frac{0.54}{\sqrt{0.0574 + 0.2528}} = \frac{0.54}{0.557} = 0.97$$

Step 4 The probability distribution:

Again, we can use either the p -value or the classical procedure:

(i) Using the p -value procedure:

a. Use both tails because H_a expresses concern for values related to “different than.”

$$P = p\text{-value} = P(t^* < -0.97) + P(t^* > 0.97) = 2 \times P(|t^*| > 0.97 | df = 9)$$

as in Figure 10.9.

To find the p -value, you have three options:

1. Use Statistical Table 4 in Appendix: $0.20 < P < 0.50$.
2. Use Statistical Table 5 in Appendix to place bounds on the p -value: $0.340 < P < 0.394$.
3. Use a computer or calculator to find the p -value: $P = 2 \times 0.1787 = 0.3574$.

For specific instructions, see *Methods* 1, 2, and 3 below.

b. The p -value is not smaller than α .

(ii) Using the classical procedure:

a. The critical region is two-tailed because H_a , expresses concern for values related to “different than”. The right-hand critical value is obtained from Statistical Table 4: $t(9, 0.05) = 1.83$, see Figure 10.10.

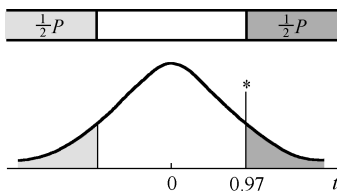


Figure 10.9 Finding p -value

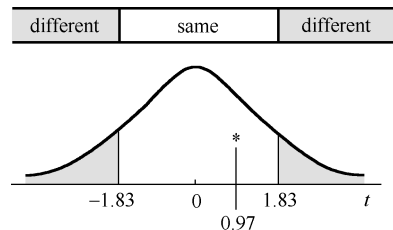


Figure 10.10 From Statistical Table 4 in Appendix

For specific instructions, see Section 9.1.

b. t^* is not in the critical region, as shown in black on Figure 10.10.

Step 5 a. Decision: Fail to reject H_o .

b. Conclusion: The evidence is not sufficient to show that machine A dispenses a different amount of soft drink than machine B, at the 0.10 level of significance. Thus, for lack of evidence we will proceed as though the two machines dispense, on average, the same amount.

To Find the p -Value Use One of Three Methods:

Method 1: Use Statistical Table 4. Find 0.97 between two entries in the $df = 9$ row and read the bounds for P from the two-tail heading at the top of the table: $0.20 < P < 0.50$.

Method 2: Use Statistical Table 5. Find $t^* = 0.97$ between two rows and $df = 9$ between two columns; read the bounds for $P(t^* > 0.97 | df = 9)$: $0.170 < \frac{1}{2}P < 0.197$; therefore, $0.340 < P < 0.394$.

Method 3: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value (do not double) for you.

New Words and Expressions

aka [ˌeɪ keɪ 'eɪ] *abbr.* 又叫做, 亦称 (also known as 的缩写)

brewer ['bruːə(r)] *n.* 啤酒制造者; 阴谋家

brewery ['bruːəri] *n.* 啤酒厂, 酿酒厂

Technical Terms

Probable Error of a Mean in *Biometrika*, 6 (1908), 1-25. 平均数的可能误差

Guinness Brewing Company 吉尼斯酿造公司

10.4 Inferences Concerning the Difference between Proportions

Using Two Independent Samples

We are often interested in making statistical comparisons between the **proportions**, **percentages**, or **probabilities** associated with two populations.

Example 10.11

A strategy for mastery of sectional anatomy must contain research to determine whether or not the methodology is sound. The research-based approach presented here demonstrates the effectiveness of a specific strategy. One application that deals with a comprehensive understanding of human anatomical features and their adjacent structures is a prescribed, *sequenced labeling method* (PSLM). The method contrasts the modern convention of random labeled human anatomical sections. The PSLM is based on studies utilizing images acquired from The Visible Human Project and was performed at Triton College. These studies have shown a significant impact on the learning rate and the comprehension of structures and adjacent anatomical relationships.

The initial component of this research consisted of a group of 28 students who were divided into two groups: Group A: 15, Group B: 13. Both groups were given a pretest, a study period, and a singular posttest. Both images presented for study were identical cross sections of the brain. The

Group A image for study was labeled in accordance with PSLM protocol. The Group B image for study was randomly labeled. Both groups were instructed to “list and recognize the parts/layers of a transverse brain section, from superficial to deep” by writing their answers in spaces provided as the only posttest in this preliminary study. Group A’s average score was 9.6667 of a total possible right of 11 with a standard deviation of 1.589. Group B scored on average 3.1538 out of 11 possible points. The mean difference between Group A and Group B is 6,5129, which is highly significant.

Table 10.6 First Set of PSLM Studies

Variable	Number of Cases	Mean	SD	SE of Mean
Group A	15	9.6667	1.589	0.410
Group B	13	3.1538	3.023	0.839

Variances	t-value df	df	2-Tail Sig	SE of Difference	95% CI of Diff
Unequal	6.98	17.57	0.000	0.933	(4.548, 8.477)

Source: Alexander Lnae, Ph D Triton College, River Grove, Illinois.

These questions ask for such comparisons: Is the proportion of homeowners who favor a certain tax proposal different from the proportion of renters who favor it? Did a larger percentage of this semester’s class than of last semester’s class pass statistics? Is the probability of a Democratic candidate winning in New York greater than the probability of a Republican candidate winning in Texas? Do students’ opinions about the new code of conduct differ from those of the faculty? You have probably asked similar questions.

In this section, we will compare two population proportions by using the difference between the observed proportions, $p'_1 - p'_2$, of two independent samples. The observed difference, $p'_1 - p'_2$, belongs to a sampling distribution with the characteristics described in the following statement.

If independent samples of sizes n_1 and n_2 are drawn randomly from large populations with $p_1 = P_1(\text{success})$ and $p_2 = P_2(\text{success})$, respectively. Then the sampling distribution of $p'_1 - p'_2$ has these properties:

(i) mean $\mu_{p'_1 - p'_2} = p_1 - p_2$,

(ii) standard error $\mu_{p'_1 - p'_2} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$, (10.10),

(iii) an approximately normal distribution if n_1 and n_2 are sufficiently large.

Recall: the properties of a binomial experiment:

1. The observed probability is $p' = x / n$, where x is the number of observed successes in n trials.
2. $q' = 1 - p'$.
3. p is the probability of success on an individual trial in a binomial probability experiment of n repeated independent trials.

The 3 “ p ” words (proportion, percentage, probability) are all the binomial parameter p , $P(\text{success})$.

In practice, we use the following **guidelines to ensure normality**:

(I) The sample sizes are both larger than 20.

(II) The products n_1p_1 , n_1q_1 , n_2p_2 , and n_2q_2 are all larger than 5.

(III) The samples consist of less than 10% of their respective populations.

Note: p_1 and P_2 are unknown; therefore, the products mentioned in guideline 2 will be estimated by $n_1p'_1, n_1q'_1, n_2p'_2$, and $n_2q'_2$.

Inferences about the difference between two population proportions, $p_1 - p_2$, will be based on the following assumptions.

Assumption for inferences about the difference between two proportions $p_1 - p_2$:

The n_1 random observations and the n_2 random observations that form the two samples are selected independently from two populations that are not changing during the sampling.

10.4.1 Confidence Interval Procedure

When we estimate the difference between two proportions, $p_1 - p_2$, we will base our estimates on the sample statistic $p'_1 - p'_2$. The point estimate, $p'_1 - p'_2$, becomes the center of the confidence interval and the confidence interval limits are found using the following formula:

Confidence Interval for the Difference between Two Proportions

$$(p'_1 - p'_2) - z(\alpha / 2) \cdot \sqrt{\left(\frac{p'_1q'_1}{n_1}\right) + \left(\frac{p'_2q'_2}{n_2}\right)} \text{ to } (p'_1 - p'_2) + z(\alpha / 2) \cdot \sqrt{\left(\frac{p'_1q'_1}{n_1}\right) + \left(\frac{p'_2q'_2}{n_2}\right)} \quad (10.10)$$

Constructing a Confidence Interval for the Difference between Two Proportions

Example 10.12

In order to look more closely at how to construct this confidence interval, let's look at sample data from a campaign plan. In studying his campaign plans, Mr. Morris wishes to estimate the difference between men's and women's views regarding his appeal as a candidate. He asks his campaign manager to take two random independent samples and find the 99% confidence interval for the difference. A sample of 1,000 voters was taken from each population, with 388 men and 459 women favoring Mr. Morris. We will use the five-step process to estimate the difference between these two proportions.

Step 1 Parameter of interest: $p_w - p_m$, the difference between the proportion of women voters and the proportion of men voters who plan to vote for Mr. Morris. Note that it is customary to place the larger value first. That way, the point estimate for the difference is a positive value.

Step 2 a. Assumptions: The samples are randomly and independently selected.

b. Probability distribution: The standard normal distribution. The populations are large (all voters); the sample sizes are larger than 20; and the estimated values for $n_m p_m, n_m q_m, n_w p_w$, and $n_w q_w$ are all larger than 5. Therefore, the sampling distribution of $p'_w - p'_m$ should have an approximately normal distribution.

c. Level of confidence: 1 $\alpha = 0.99$.

Step 3 Sample information:

We have $n_m = 1,000$, $x_m = 388$, $n_w = 1,000$, and $x_w = 459$.

$$p'_m = \frac{x_m}{n_m} = \frac{388}{1000} = 0.388 \quad q'_m = 1 - 0.388 = 0.612$$

$$p'_w = \frac{x_w}{n_w} = \frac{459}{1000} = 0.459 \quad q'_w = 1 - 0.459 = 0.541$$

Step 4 a. Confidence coefficient: This is a two-tailed situation, with $\alpha/2$ in each tail. From Statistical Table 2(II), $z(\alpha/2) = z(0.005) = 2.58$, see Figure 10.11. Instructions for using Statistical Table 2(II) are in Appendix.

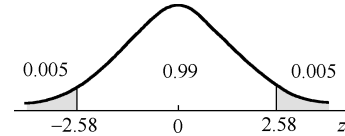


Figure 10.11 From Statistical Table 2 (II)

b. Maximum error of estimate: Using the maximum error part of formula (10.10), we have

$$E = z(\alpha/2) \cdot \sqrt{\left(\frac{p'_w q'_w}{n_w}\right) + \left(\frac{p'_m q'_m}{n_m}\right)}$$

$$E = 2.58 \cdot \sqrt{\left(\frac{(0.459)(0.541)}{1000}\right) + \left(\frac{(0.388)(0.612)}{1000}\right)}$$

$$= 2.58 \sqrt{0.000248 + 0.000237} = (2.58)(0.022) = 0.057$$

c. Lower and upper confidence limits:

$$(p'_w - p'_m) \pm E$$

$$0.071 \pm 0.057$$

$$0.071 - 0.057 = 0.014 \quad \text{to} \quad 0.071 + 0.057 = 0.128$$

Step 5 Confidence interval: 0.014 to 0.128 is the 99% confidence interval for $p_w - p_m$. With 99% confidence, we can say that there is a difference of from 1.4% to 12.8% in Mr. Morris's voter appeal. That is, a larger proportion of women than men favor Mr. Morris, and the difference in the proportions is between 1.4% and 12.8%.

Confidence intervals and hypothesis tests can sometimes be interchanged; that is, a confidence interval can be used in place of a hypothesis test. The previous example with Mr. Morris called for a confidence interval. Now suppose that Mr. Morris asked, "Is there a difference in my voter appeal to men voters as opposed to women voters?"

To answer his question, you would not need to complete a hypothesis test if you chose to test at $\alpha = 0.01$ using a two-tailed test. "No difference" would mean a difference of zero, which is not included in the interval from 0.014 to 0.128 (the interval determined in the example). Therefore, a null hypothesis of "no difference" would be rejected, thereby substantiating the conclusion that a significant difference exists in voter appeal between the two groups.

10.4.2 Hypothesis-Testing Procedure

When we test the null hypothesis "There is no difference between two proportions", the test

statistic will be the difference between the observed proportions divided by the **standard error**; it is found with the following formula:

Test Statistic for the Difference between Two Proportions—Population Proportion Known

$$z^* = \frac{p'_1 - p'_2}{\sqrt{pq \left[\left(\frac{1}{n_1} \right) + \left(\frac{1}{n_2} \right) \right]}} \quad (10.11)$$

Notes:

1. The null hypothesis is $p_1 = p_2$, or $p_1 - p_2 = 0$ (the difference is zero).
2. Nonzero differences between proportions are not discussed in this section.
3. The numerator of formula (10.12) could be written as $(p'_1 - p'_2) - (p_1 - p_2)$ but since the null hypothesis is assumed to be true during the test, $p_1 - p_2 = 0$. By substitution, the numerator becomes simply $p'_1 - p'_2$.

4. Since the null hypothesis is $p_1 = p_2$, the standard error of $p'_1 - p'_2$, $\sqrt{\left(\frac{p_1 q_1}{n_1} \right) + \left(\frac{p_2 q_2}{n_2} \right)}$, can be written as $\sqrt{pq \left[\left(\frac{1}{n_1} \right) + \left(\frac{1}{n_2} \right) \right]}$, where $p = p_1 = p_2$ and $q = 1 - p$.

5. When the null hypothesis states that $p_1 = p_2$ and does not specify the value of either p_1 or p_2 , the two sets of sample data will be pooled to obtain the estimate for p . This pooled probability (known as p'_p) is the total number of successes divided by the total number of observations with the two samples combined; it is found using the next formula:

$$p'_p = \frac{x_1 + x_2}{n_1 + n_2} \quad (10.12)$$

and q'_p is its complement,

$$q'_p = 1 - p'_p \quad (10.13)$$

When the pooled estimate, p'_p , is being used formula (10.12) becomes formula (10.15):

Test Statistic for the Difference between Two Proportions--Population Proportion Unknown

$$z^* = \frac{p'_1 - p'_2}{\sqrt{(p'_p)(q'_p) \left[\left(\frac{1}{n_1} \right) + \left(\frac{1}{n_2} \right) \right]}} \quad (10.14)$$

One-Tailed Hypothesis Test for the Difference between Two Proportions

Example 10.13

To examine the difference between two proportions more closely, think of a salesman for a new manufacturer of cellular phones. He claims not only that they cost the retailer less but also that the percentage of defective cellular phones found among his products will be no higher than the percentage of defectives found in a competitor's line. To test this statement, the retailer took random samples of

each manufacturer's product. The sample summaries are given in Table 10.7. Can we reject the salesman's claim at the 0.05 level of significance? Let's use the five-step process to find out.

Table 10.7 Cellular Phone Sample Information

Product	Number Defective	Number Checked
Salesman's	15	150
Competitor's	6	150

Step 1 a. Population parameter of interest: $p_s - p_c$, the difference between the proportion of defectives in the salesman's product and the proportion of defectives in the competitor's product.

b. Statement of hypotheses: The concern of the retailer is that the salesman's less expensive product may be of a poorer quality, meaning a greater proportion of defectives. If we use the difference "suspected larger proportion - smaller proportion", then the alternative hypothesis is "The difference is positive (greater than zero)."

$H_o: p_s - p_c = 0$ (\leq) (salesman's defective rate is no higher than competitor's)

$H_a: p_s - p_c > 0$ (salesman's defective rate is higher than competitor's)

Step 2 a. Assumptions: Random samples were selected from the products of two different manufacturers.

b. Probability distribution: The standard normal distribution. Populations are very large (all cellular phones produced); the samples are larger than 20; and the estimated products $n_s p'_s, n_s q'_s, n_c p'_c$, and $n_c q'_c$ are all larger than 5. Therefore, the sampling distribution should have an approximately normal distribution, z^* will be calculated using formula (10.15).

c. Determine the level of significance: $\alpha = 0.05$.

Step 3 a. Sample information:

$$p'_s = \frac{s_s}{n_s} = \frac{15}{150} = 0.10$$

$$p'_c = \frac{s_c}{n_c} = \frac{6}{150} = 0.04$$

$$p'_p = \frac{x_1 + x_2}{n_1 + n_2} = \frac{15 + 6}{150 + 150} = \frac{21}{300} = 0.07$$

$$q'_p = 1 - p'_p = 1 - 0.07 = 0.93$$

b. Test statistic

$$z^* = \frac{p'_1 - p'_2}{\sqrt{(p'_p)(q'_p) \left[\left(\frac{1}{n_s} \right) + \left(\frac{1}{n_c} \right) \right]}}$$

$$z^* = \frac{0.10 - 0.04}{\sqrt{(0.07)(0.93) \left[\left(\frac{1}{150} \right) + \left(\frac{1}{150} \right) \right]}} = \frac{0.06}{\sqrt{0.000868}} = \frac{0.06}{0.02946} = 2.04$$

Step 4 The probability distribution:

Again, we can use either the p -value or the classical procedure:

(i) Using the p -value procedure:

a. Use the right-hand tail because H_a expresses concern for values related to “higher than”. $P = p\text{-value} = P(z^* > 2.04)$ as shown in Figure 10.12.

To find the p -value, you have three options:

1. Use Statistical Table 1 in Appendix to calculate the p -value: $P = 0.5000 - 0.4793 = 0.0207$.
2. Use Statistical Table 3 in Appendix to place bounds on the p -value: $0.0202 < P < 0.0228$.
3. Use a computer or calculator: $P = 0.0207$. For specific instructions, see Section 8.5.

b. The p -value is smaller than α .

(ii) Using the classical procedure:

a. The critical region is the right-hand tail because H_a expresses concern for values related to “higher than”. The critical value is obtained from Statistical Table 2(I) $z(0.05) = 1.65$, see Figure 10.13.

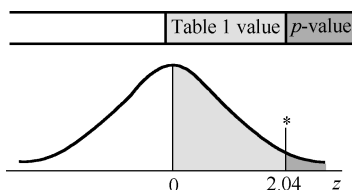


Figure 10.12 From Statistical Table 1

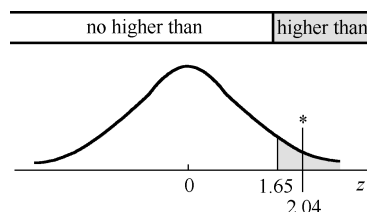


Figure 10.13 From Statistical Table 2 (II)

For specific instructions, see Section 8.6.

b. z^* is in the critical region, as shown in black on Figure 10.13.

Step 5 The results:

a. **Decision:** Reject H_0 .

b. **Conclusion:** At the 0.05 level of significance, there is sufficient evidence to reject the salesman's claim; the proportion of his company's cellular phones that are defective is higher than the proportion of his competitor's cellular phones that are defective.

New Words and Expressions

mastery ['mɑ:stəri] *n.* 精通, 熟练; 统治, 控制; 优势

anatomy [ə'nætəmi] *n.* 解剖, 分解, 分析; (详细的) 剖析; (生物体) 解剖结构

anatomical [ˌænə'tɒmɪkl] *adj.* 结构上的, 解剖的

effectiveness [ɪ'fek'tɪvnis] *n.* 有效, 有力; 有效性; 效益

protocol ['prəʊtəkol] *n.* 礼仪; (数据传递) 协议; 科学实验报告 (或计划)

transverse ['trænzvɜ:s] *adj.* 横向的; 横断的; 向横活动的 *n.* 横向物; 横轴; 横断面

homeowner ['həʊməʊnə(r)] *n.* 自己拥有住房者, (住自己房子的) 私房屋主

cellular ['seljələ(r)] *adj.* 细胞的; 由细胞组成的; 多孔的

retailer ['ri:tɪlə(r)] *n.* 零售商, 零售店; 传播的人, 到处散布闲话的人
competitor [kəm'petɪtə(r)] *n.* 竞争者; 对手

Technical Terms

Democratic candidate 民主党候选人
Republican candidate 共和党候选人
Triton College 特里顿学院
transverse brain 脑横截面

10.5 Inferences Concerning the Ratio of Variances Using Two Independent Samples

When comparing two populations, we naturally compare their two most fundamental distribution characteristics, their “center” and their “spread”, by comparing their means and standard deviations.

We have learned, in two of the previous sections, how to use the *t*-distribution to make inferences comparing two population means with either dependent or independent samples. These procedures were intended to be used with normal populations, but they work quite well even when the populations are not exactly normally distributed.

The next logical step in comparing two populations is to compare their standard deviations, the most often used measure of spread. However, sampling distributions that deal with sample standard deviations (or variances) are very sensitive to slight departures from the assumptions. Therefore, the only inference procedure to be presented here will be the **hypothesis test for the equality of standard deviations (or variances)** for two normal populations.

The soft-drink bottling company discussed in Section 9.3 is trying to decide whether to install a modern, high-speed bottling machine. There are, of course, many concerns in making this decision, and one of them is that the increased speed may result in increased variability in the amount of fill placed in each bottle; such an increase would not be acceptable. To this concern, the manufacturer of the new system responded that the variance in fills will be no greater with the new machine than with the old. (The new system will fill several bottles in the same amount of time as the old system fills one bottle; this is the reason the change is being considered.) A test is set up to statistically test the bottling company’s concern, “Standard deviation of new machine is greater than standard deviation of old,” against the manufacturer’s claim, “Standard deviation of new is no greater than standard deviation of old.”

10.5.1 Writing for the Equality of Variances

Imagine you were given the task of stating the null and alternative hypotheses to be used for comparing the variances of the two soft-drink bottling machines. You could do this using one of several

equivalent ways to express the null and alternative hypotheses, but since the test procedure uses the ratio of variances, the recommended convention is to express the null and alternative hypotheses as ratios of the population variances. Furthermore, it is recommended that the “larger” or “expected to be larger” variance be the numerator. The concern of the soft-drink company is that the new modern machine (m) will result in a larger standard deviation in the amounts of fill than its present machine (p); $\sigma_m > \sigma_p$, or equivalently $\sigma_m^2 > \sigma_p^2$, which becomes $\frac{\sigma_m^2}{\sigma_p^2} > 1$. We want to test the manufacturer’s claim (the null hypothesis) against the company’s concern (the alternative hypothesis):

$$H_o: \frac{\sigma_m^2}{\sigma_p^2} = 1 (\leq) \text{ (} m \text{ is no more variable)}$$

$$H_o: \frac{\sigma_m^2}{\sigma_p^2} > 1 \text{ (} m \text{ is more variable)}$$

10.5.2 Using the F -Distribution

Inferences about the ratio of variances for two normally distributed populations use the F -distribution. The F -distribution, similar to the Student’s t -distribution and the χ^2 -distribution, is a family of probability distributions. Each F -distribution is identified by two numbers of degrees of freedom, one for each of the two samples involved.

Before continuing with the details of the hypothesis-testing procedure, let’s learn about the F -distribution.

Properties of the F-distribution:

- (i) F is nonnegative; it is zero or positive.
- (ii) F is nonsymmetrical; it is skewed to the right.
- (iii) F is distributed so as to form a family of distributions; there is a separate distribution for each pair of numbers of degrees of freedom.

For inferences discussed in this section, the number of degrees of freedom for each sample is $df_1 = n_1 - 1$ and $df_2 = n_2 - 1$. Each different combination of degrees of freedom results in a different F -distribution, and each F -distribution looks approximately like the distribution shown in Figure 10.14.

The critical values for the F -distribution are identified using three values:

- (I) df_n the degrees of freedom associated with the sample whose variance is in the numerator of the calculated F ,
- (II) df_d , the degrees of freedom associated with the sample whose variance is in the denominator, and
- (III) α the area under the distribution curve to the right of the critical value being sought.

Therefore, the symbolic name for a critical value of F will be $F(df_n, df_d, \alpha)$, as shown in Figure 10.15.

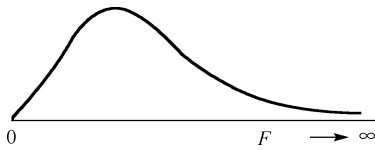


Figure 10.14 F-Distribution

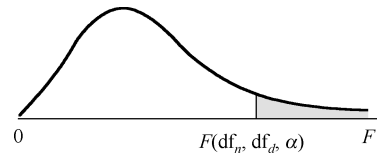


Figure 10.15 A critical value of F

Since it takes three values to identify a single critical value of F , making tables for F is not as simple as with previously studied distributions. The tables presented in this textbook are organized so as to have a different table for each different value of α , the “area to the right”. Statistical Table 7(I) in Appendix shows the critical values for $F(df_n, df_d, \alpha)$ when $\alpha = 0.05$; Statistical Table 7(II) gives the critical values when $\alpha = 0.025$; Table 7(III) gives the values when $\alpha = 0.01$.

If we wanted to find $F(5, 8, 0.05)$, the critical F -value for samples of size 6 and size 9 with 5% of the area in the right-hand tail, we would need to consult Statistical Table 7(I) ($\alpha = 0.05$). Using the partial view of Statistical Table 7(I) below, notice that the intersection of column $df = 5$ (for the numerator) and row $df = 8$ (for the denominator) occurs at the value: $F(5, 8, 0.05) = 3.69$, see Table 10.8.

Table 10.8 Portion of Statistical Table 7(I)

Portion of Statistical Table 7(I) ($\alpha = 0.05$)					
df for Denominator	df for Numerator				
	...	5	...	8	...
	⋮				
	5			4.82	
	⋮				
	8		3.69		
	⋮				

$F(8, 5, 0.05) = 4.82$
 $F(5, 8, 0.05) = 3.69$

You can also see that if the two degrees of freedom are reversed, the resulting F is different: $F(8, 5, 0.05)$ is 4.82. The degrees of freedom associated with the numerator and with the denominator *must* be kept in the correct order; 3.69 is different from 4.82. Check some other pairs to verify that interchanging the degrees of freedom numbers will result in different F -values.

Use of the F -distribution has a condition. That is, we make certain assumptions for inferences about the ratio of two variances: (1) the samples are randomly selected from normally distributed populations, and (2) the two samples are selected in an independent manner.

10.5.3 One-Tailed Hypothesis Test for the Equality of Variances

Test Statistic for Equality of Variances

$$F^* = \frac{s_n^2}{s_d^2}, \quad \text{with } df = n_n - 1 \text{ and } df_d = n_d - 1 \quad (10.15)$$

The sample variances are assigned to the numerator and denominator in the order established by the null and alternative hypotheses for one-tailed tests. The calculated ratio, F^* , will have an F -distribution with $df_n = n_n - 1$ (numerator) and $df_d = n_d - 1$ (denominator) when the assumptions are met and the null hypothesis is true.

Example 10.14 (Example 10.13 continued)

Recall that our soft-drink bottling company was to make a decision about the equality of the variances of amounts of fill between its present machine and a modern high-speed outfit. Does the sample information in Table 10.9 present sufficient evidence to reject the null hypothesis (the manufacturer's claim) that the modern high-speed bottle-filling machine fills bottles with no greater variance than the company's present machine? We will assume that the amounts of fill are normally distributed for both machines, and complete the five-step process using $\alpha = 0.01$.

Table 10.9 Sample Information on Variances of Fills

Sample	<i>N</i>	<i>s</i> ²
Present machine (<i>p</i>)	22	0.0008
Modern high-speed machine(<i>m</i>)	25	0.0018

Step 1 a. Parameter of interest: $\frac{\sigma_m^2}{\sigma_p^2}$ the ratio of the variances in the amounts of fill

placed in bottles for the modern machine versus the company's present machine.

b. Statement of hypotheses: The hypotheses were established at the beginning of this section on page 229:

$$H_o: \frac{\sigma_m^2}{\sigma_p^2} = 1 (\leq) \text{ (} m \text{ is no more variable)}$$

$$H_o: \frac{\sigma_m^2}{\sigma_p^2} > 1 \text{ (} m \text{ is more variable)}$$

Note: When the “expected to be larger” variance is in the numerator for a one-tailed test, the alternative hypothesis states “The ratio of the variances is greater than one.”

Step 2 a. Assumptions: The sampled populations are normally distributed (given in the statement of the problem), and the samples are independently selected (drawn from two separate populations).

b. Test statistic: The *F*-distribution with the ratio of the sample variances and formula (10.16).

c. Level of significance: $\alpha = 0.01$.

Step 3 a. Sample information: See Table 10.8.

b. Calculated test statistic: Using formula (10.16), we have

$$F^* = \frac{s_m^2}{s_p^2} : F^* = \frac{0.0018}{0.0008} = 2.25$$

The number of degrees of freedom for the numerator is $df_n = 24$ (or $25 - 1$) because the sample from the modern high-speed machine is associated with the numerator, as specified by the null hypothesis. Also, $df_d = 21$ because the sample associated with the denominator has size 22.

Step 4 The probability distribution:

Again, we can use either the *p*-value or the classical procedure:

(i) Using the p -value procedure:

a. Use the right-hand tail because H_a expresses concern for values related to “more than”. $P = P(F^* > 2.25, \text{ with } df_n = 24 \text{ and } df_d = 21)$ as shown on Figure 10.16.

To find the p -value, you have two options:

1. Use Statistical Tables 7(I) and Statistical Table 7(II) in Appendix to place bounds on the p -value: $0.025 < P < 0.05$.

2. Use a computer or calculator to find the p -value: $P = 0.0323$.

Specific instructions follow this example.

b. The p -value is not smaller than the level of significance, α (0.01).

(ii) Using the classical procedure:

a. The critical region is the right-hand tail because H_a expresses concern for values related to “more than”. $df_n = 24$ and $df_d = 21$. The critical value is obtained from Statistical Table 7(III): $F(24, 21, 0.01) = 2.80$.

For additional instructions, see Section 10.5 ahead.

b. F^* is not in the critical region, as shown in black on Figure 10.17.

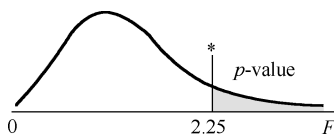


Figure 10.16 Finding the p -value procedure

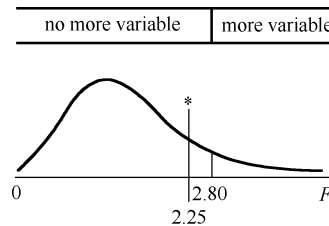


Figure 10.17 From Statistical Table 7(III)

Step 5 The results

a. Decision: Fail to reject H_0 :

b. Conclusion: At the 0.01 level of significance, the samples do not present sufficient evidence to indicate an increase in variance.

Calculating the p -Value When Using the F -Distribution

There are two methods for calculating the p -value when using the F -distribution:

Method 1: Use Statistical Table 7 in Appendix to place bounds on the p -value. Using Statistical Tables 7(I), 7(II), and 7(III) in Appendix to estimate the p -value is very limited. However, for the softdrink example we just worked through comparing variances of fill amounts for the present machine and a more modern machine, the p -value can be estimated. By inspecting Statistical Tables 7(I) and 7(II), you will find that $F(24, 21, 0.025) = 2.37$ and $F(24, 21, 0.05) = 2.05$. $F^* = 2.25$ is between the values 2.37 and 2.05; therefore, the p -value is between 0.025 and 0.05: $0.025 < P < 0.05$. See Figure 10.18 to the right.

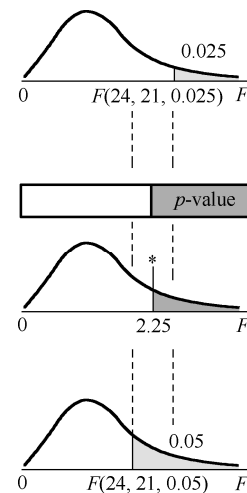


Figure 10.18 From Statistical Table 7(I), 7(II), and 7(III)

Method 2: If you are doing the hypothesis test with the aid of a computer or calculator, most likely it will calculate the p -value for you.

10.5.4 Critical F -Values for One- and Two-Tailed Tests

The tables of critical values for the F -distribution give only the right-hand critical values. This will not be a problem because the right-hand critical value is the only critical value that will be needed. You can adjust the numerator-denominator order so that all the “activity” is in the right-hand tail. There are two cases: one-tailed tests and two-tailed tests.

One-tailed tests: Arrange the null and alternative hypotheses so that the alternative is always “greater than”. The F^* -value is calculated using the same order as specified in the null hypothesis (recall the soft drink bottling example).

Two-tailed tests: When the value of F , is calculated, always use the sample with the larger variance for the numerator; this will make F^* greater than one and place it in the right-hand tail of the distribution. Thus, you will need only the critical value for the right-hand tail.

All hypothesis tests about two variances can be formulated and completed in a way that both the critical value of F and the calculated value of F^* will be in the right-hand tail of the distribution. Since Statistical Tables 7(I), 7(II), and 7(III) contain only critical values for the right-hand tail, this will be convenient and you will never need critical values for the left-hand tail.

Example 10.15 Format for Writing Hypotheses for the Equality of Variances

Reorganize the alternative hypothesis so that the critical region will be the right-hand tail:

$$H_o : \sigma_1^2 < \sigma_2^2 \text{ or } \frac{\sigma_1^2}{\sigma_2^2} < 1 \quad (\text{population 1 is less variable})$$

Solution

Reverse the direction of the inequality, and reverse the roles of the numerator and denominator.

$$H_o : \sigma_2^2 > \sigma_1^2 \text{ or } \frac{\sigma_2^2}{\sigma_1^2} > 1 \quad (\text{population 2 is more variable})$$

The calculated test statistic F^* will be $\frac{s_2^2}{s_1^2}$.

Example 10.16 Two-Tailed Hypotheses Test for the Equality of Variances

Find F^* and the critical values for the following hypothesis test so that only the right-hand critical value is needed. Use $\alpha = 0.05$ and the sample information $n_1 = 10$, $n_2 = 8$, $s_1 = 5.4$ and $s_2 = 3.8$.

$$H_o : \sigma_1^2 = \sigma_2^2 \text{ or } \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a : \sigma_1^2 \neq \sigma_2^2 \text{ or } \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Solution

When the alternative hypothesis is two-tailed (\neq), the calculated F^* can be either $F^* = \frac{s_1^2}{s_2^2}$ or $F^* = \frac{s_2^2}{s_1^2}$. The choice is ours; we need only make sure that we keep df_n and df_d in the correct order.

We make the choice by looking at the sample information and using the sample with the larger standard deviation or variance as the numerator. Therefore, in this illustration,

$$F^* = \frac{s_1^2}{s_2^2} = \frac{5.4^2}{3.8^2} = \frac{29.16}{14.44} = 2.02$$

The critical values for this test are left tail, $F(9, 7, 0.975)$, and right tail, $F(9, 7, 0.025)$, as shown in Figure 10.19.

Since we choose the sample with larger standard deviation (or variance) for the numerator, the value of F^* will be greater than 1 and will be in the right-hand tail; therefore, only the right-hand critical value is needed. All critical values for left-hand tails will be values between 0 and 1.

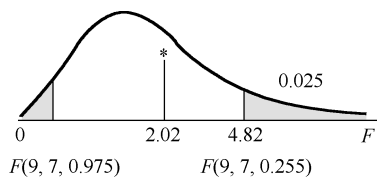


Figure 10.19 From Statistical Tables 7

New Words and Expressions

soft-drink [sɒft drɪŋk] *n.* 软饮料 (不含酒精的饮料)

bottle ['bɒtl] *n.* 瓶子; 一瓶 (的量); (婴儿) 奶瓶 *vt.* 把.....装入瓶中; 控制

reverse [rɪ'vɜ:s] *vt.* 彻底转变; 使完全相反; 撤销, 废除 (决定、法律等); 使次序颠倒
n. 倒转, 反向; 倒退; 失败

Summary

In this unit we began the comparisons of two populations by distinguishing between independent and dependent samples, which are statistically important and useful sampling procedures. We then proceeded to examine the inferences concerning the comparison of means, proportions, and variances for two populations.

We are always making comparisons between two groups. We compare means and we compare proportions. In this unit we have learned how to statistically compare two populations by making inferences about their means, proportions, or variances.

In Units 8, 9, and 10 we have learned how to use confidence intervals and hypothesis tests to answer questions about means, proportions, and standard deviations for one or two populations. From here we can expand our techniques to include inferences about more than two populations as well as inferences of different types.

Problems

10.1 The students at a local high school were assigned to do a project for their statistics class. The project involved having sophomores take a timed test on geometric concepts. The statistics students would then use these data to determine whether there was a difference between male and female performances. Would the resulting sets of data represent dependent or independent samples? Explain.

10.2 Twenty people were selected to participate in a psychology experiment. They answered a short multiple-choice quiz about their attitudes on a particular subject and then viewed a 45-minute film. The following day the same 20 people were asked to answer a follow-up questionnaire about their attitudes. At the completion of the experiment, the experimenter will have two sets of scores. Do these two samples represent dependent or independent samples? Explain.

10.3 A study is being designed to determine the reasons why adults choose to follow a healthy diet plan. The study will survey 1000 men and 1000 women. Upon completion of the study, the reasons men choose a healthy diet will be compared with the reasons women choose a healthy diet.

- How can the data be collected if independent samples are to be obtained? Explain in detail.
- How can the data be collected if dependent samples are to be obtained? Explain in detail.

10.4 Given this set of paired data:

Pairs	1	2	3	4	5
Sample A	3	6	1	4	7
Sample B	2	5	1	2	8

Find:

- The paired differences, $d = A - B$, for this set of data
- The mean \bar{d} of the paired differences
- The standard deviation s_d of the paired differences

10.5 Salt-free diets are often prescribed to people with high blood pressure. The following data values were obtained from an experiment designed to estimate the reduction in diastolic blood pressure as a result of consuming a salt-free diet for 2 weeks. Assume diastolic readings to be normally distributed.

Before	93	106	87	92	102	95	88	110
After	92	102	89	92	101	96	88	105

a. What is the point estimate for the mean reduction in the diastolic reading after 2 weeks on this diet?

- Find the 98% confidence interval for the mean reduction.

10.6 Determine the p -value for each hypothesis test for the mean difference.

- $H_o: \mu_d = 0$ and $H_a: \mu_d > 0$, with $n = 20$ and $t^* = 1.86$
- $H_o: \mu_d = 0$ and $H_a: \mu_d < 0$, with $n = 20$ and $t^* = -1.86$
- $H_o: \mu_d = 0$ and $H_a: \mu_d < 0$, with $n = 29$ and $t^* = -2.63$
- $H_o: \mu_d = 0.75$ and $H_a: \mu_d > 0.75$, with $n = 10$ and $t^* = 3.57$

10.7 Ten randomly selected college students, who participated in a learning community,

were given pre-self-esteem and post-self-esteem surveys. A learning community is a group of students who take two or more courses together. Typically, each learning community has a theme, and the faculty involved coordinate assignments linking the courses. Research has shown that the benefits of higher self-esteem, higher grade point averages (GPAs), and improved satisfaction in courses, as well as better retention rates, result from involvement in a learning community. The scores on the surveys are as follows:

Student	1	2	3	4	5	6	7	8	9	10
Prescore	18	14	11	23	19	21	21	21	11	22
Postscore	17	17	10	25	20	10	24	22	10	24

Does this sample of students show sufficient evidence that self-esteem scores were higher after participation in a learning community? Lower scores indicate higher self-esteem. Use the 0.05 level of significance and assume normality of scores.

10.8 To test the effect of a physical fitness course on one's physical ability, the number of sit-ups that a person could do in 1 minute, both before and after the course, was recorded. Ten randomly selected participants scored as shown in the following table, Can you conclude that a significant amount of improvement took place? Use $\alpha = 0.01$ and assume normality.

Before	29	22	25	29	26	24	31	46	34	28
After	30	26	25	35	33	36	32	54	50	43

- Solve using the t-value approach.
- Solve using the classical approach.

10.9 A study comparing attitudes toward death was conducted in which organ donors (individuals who had signed organ donor cards) were compared with nondonors. The study is reported in the journal *Death Studies*. Templer's Death Anxiety Scale (DAS) was administered to both groups. On this scale, high scores indicate high anxiety concerning death. The results were reported as follows.

	n	Mean	Std.Dev.
Organ Donors	25	5.36	2.91
Nonorgan Donors	69	7.62	3.45

Construct the 95% confidence interval for the difference between the means, $\mu_{\text{non}} - \mu_{\text{donor}}$.

10.10 At a large university, a mathematics placement exam is administered to all students. Samples of 36 male and 30 female students are randomly selected from this year's student body and the following scores recorded:

Male	72	68	75	82	81	60	75	85	80	70
	71	84	68	85	82	80	54	81	86	79
	99	90	68	82	60	63	67	72	77	51
	61	71	81	74	79	76				
Female	81	76	94	89	83	78	85	91	83	83
	84	80	84	88	77	74	63	69	80	82
	89	69	74	97	73	79	55	76	78	81

Construct the 95% confidence interval for the difference between the mean scores for male and female students.

10.11 If a random sample of 18 homes south of Center Street in Provo has a mean selling price of \$145,200 and a standard deviation of \$4,700, and a random sample of 18 homes north of Center Street has a mean selling price of \$148,600 and a standard deviation of \$5,800, can you conclude that there is a significant difference between the selling price of homes in these two areas of Provo at the 0.05 level? Assume normality.

- Solve using the p -value approach.
- Solve using the classical approach.

10.12 Twenty laboratory mice were randomly divided into two groups of 10. Each group was fed according to a prescribed diet. At the end of 3 weeks, the weight gained by each animal was recorded. Do the data in the following table justify the conclusion that the mean weight gained on diet B was greater than the mean weight gained on diet A, at the $\alpha = 0.05$ level of significance? Assume normality.

Diet A	5	14	7	9	11	7	13	14	12	8
Diet B	5	21	16	23	4	16	13	19	9	21

10.13 If $n_1 = 40$, $p_1' = 0.9$, $n_2 = 50$, and $p_2' = 0.9$:

- Find the estimated values for both np 's and both nq 's.
- Would this situation satisfy the guidelines for approximately normal? Explain.

10.14 Calculate the estimate for the standard error of the difference between two proportions for each of the following cases:

- $n_1 = 40$, $p_1' = 0.8$, $n_2 = 50$, and $p_2' = 0.8$
- $n_1 = 33$, $p_1' = 0.6$, $n_2 = 38$, and $p_2' = 0.65$

10.15 The proportions of defective parts produced by two machines were compared, and the following data were collected:

Machine 1: $n = 150$; number of defective parts = 12

Machine 2: $n = 150$; number of defective parts = 6

Determine a 90% confidence interval for $p_1 - p_2$.

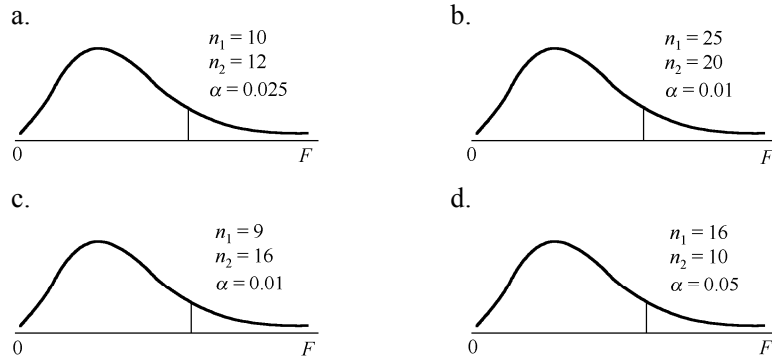
10.16 In a survey of families in which both parents work, one of the questions asked was, "Have you refused a job, promotion, or transfer because it would mean less time with your family?" A total of 200 men and 200 women were asked this question. "Yes" was the response given by 29% of the men and 24% of the women. Based on this survey, can we conclude that there is a difference in the proportion of men and women responding "yes" at the 0.05 level of significance?

10.17 State the null hypothesis, H_0 , and the alternative hypothesis, H_a , that would be used to test the following claims:

- The variances of populations A and B are not equal.
- The standard deviation of population I is larger than the standard deviation of population II.

- c. The ratio of the variances for populations A and B is different from 1.
- d. The variability within population C is less than the variability within population D.

10.18 Using the F (df_1 , df_2 , α) notation, name each of the critical values shown on the following figures.



10.19 Determine the p -value that would be used to test the following hypotheses when F is used as the test statistic:

- a. $H_o: \sigma_1 = \sigma_2$ vs. $H_a: \sigma_1 > \sigma_2$, with $n_1 = 10$, $n_2 = 16$, and $F^* = 2.47$
- b. $H_o: \sigma_1^2 = \sigma_2^2$ vs. $H_a: \sigma_1^2 > \sigma_2^2$, with $n_1 = 25$, $n_2 = 21$, and $F^* = 2.31$
- c. $H_o: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs. $H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$, with $n_1 = 41$, $n_2 = 61$, and $F^* = 4.78$
- d. $H_o: \sigma_1 = \sigma_2$ vs. $H_a: \sigma_1 < \sigma_2$, with $n_1 = 10$, $n_2 = 16$, and $F^* = 2.47$

10.20 Determine the critical region and critical value(s) that would be used to test the following hypotheses using the classical approach when F is used as the test statistic.

- a. $H_o: \sigma_1^2 = \sigma_2^2$ vs. $H_a: \sigma_1^2 > \sigma_2^2$, with $n_1 = 10$, $n_2 = 16$, and $\alpha = 0.05$
- b. $H_o: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs. $H_a: \frac{\sigma_1^2}{\sigma_2^2} \neq 1$, with $n_1 = 25$, $n_2 = 31$, and $\alpha = 0.05$
- c. $H_o: \frac{\sigma_1^2}{\sigma_2^2} = 1$ vs. $H_a: \frac{\sigma_1^2}{\sigma_2^2} > 1$, with $n_1 = 10$, $n_2 = 10$, and $\alpha = 0.01$
- d. $H_o: \sigma_1 = \sigma_2$ vs. $H_a: \sigma_1 < \sigma_2$, with $n_1 = 25$, $n_2 = 16$, and $\alpha = 0.01$

10.21 A bakery is considering buying one of two gas ovens. The bakery requires that the temperature remain constant during a baking operation. A study was conducted to measure the variance in temperature of the ovens during the baking process. The variance in temperature before the thermostat restarted the flame for the Monarch oven was 2.4 for 16 measurements. The variance for the Kraft oven was 3.2 for 12 measurements. Does this information provide sufficient reason to conclude that there is a difference in the variances for the two ovens? Assume measurements are normally distributed and use a 0.02 level of significance.

10.22 The quality of the end product is somewhat determined by the quality of the materials used. Textile mills monitor the tensile strength of the fibers used in weaving their yard goods. The following independent random samples are tensile strengths of cotton fibers from two suppliers.

Supplier A	78	82	85	83	77	84	90	82	93	82
	80	82	77	80	80					
Supplier B	76	79	83	78	72	73	69	80	74	77
	78	78	73	76	78	79				

Calculate the observed value of F , F^* , for comparing the variances of these two sets of data.

10.23 Americans snooze on the weekends, according to a poll of 1506 adults for the National Sleep Foundation and reported in a USA Today Snapshot during April 2005.

Hours of Sleep	Weekdays	Weekends
Less than 6	0.16	0.10
6-6.9	0.24	0.15
7-7.9	0.31	0.24
8 or more	0.26	0.49

Two independent random samples were taken at a large industrial complex. The workers selected in one sample were asked, “How many hours, to the nearest 15 minutes, did you sleep on Tuesday night this week?” The workers selected for the second sample were asked, “How many hours, to nearest 15 minutes, did you sleep on Saturday night last weekend?”

Weekday			Weekend			
5.00	7.75	7.25	9.00	7.25	8.75	7.50
9.25	7.25	8.75	6.25	5.25	9.25	9.25
7.00	7.75	6.75	7.50	8.50	8.75	6.50
9.25	7.00	7.75	8.00	8.75	9.50	8.00
9.25	9.25	6.00	8.75	7.75	8.75	7.50

- Construct a histogram and find the mean and standard deviation for each set of data.
- Do the distributions of “hours of sleep on weekday” and “hours of sleep on weekend” resulting from the poll appear to be similar in shape? center? spread? Discuss your responses.
- Is it possible that both of the samples were drawn from normal populations? Justify your answer.
- Is the mean number of hours slept on the weekend statistically greater than the mean number of hours slept on the weekday? Use $\alpha = 0.05$.
- Is there sufficient evidence to show that the standard deviation of these two samples are statistically different? Use $\alpha = 0.05$.
- Explain how the answers to parts b-e now affect your thoughts about your answer to part a.

Statistics is the art of making numerical conjectures about puzzling questions

— Freedman et al. 1978



Unit 11

An Introduction to Simple Regression



11.1 Regression as a Best Fitting Line



11.2 Interpreting OLS Estimates



11.3 Fitted Values and R^2 : Measuring the Fit of a Regression Model



11.4 Nonlinearity in Regression



Reading English Materials



Problems

11.1 Regression as a Best Fitting Line

Regression is the most important tool applied statistics use to understand the relationship among two or more variables. It is particularly useful for the common case where there are many variables (e.g. unemployment and interest rates, the money supply, exchange rates, inflation, etc.) and the interactions between them are complex.

To give an example, in the summer of 2008 a great deal of attention in the UK media focussed on the proper level at which interest rates should be set. In particular, the manufacturing sector complained that interest rates were too high. They argued that high interest rates encouraged foreigners to invest their money in the UK which, in turn, caused the pound to appreciate. A higher pound made it difficult for UK firms to export their products, resulting in falling sales, increased layoffs and rising unemployment.

But this is only part of the story. Still others believed that interest rates were too low, and argued that higher interest rates were necessary to choke off inflationary pressures due to a relationship between inflation and interest rates. Thus, an important economic question (i.e. interest rate determination) was at stake, and a large number of variables-interest rates, exchange rates, inflation, manufacturing output, exports, unemployment-must be considered in arriving at an answer to the problem.

All these variables (and more) shaped the discussion of what the relevant interest rate should be. As a second example, consider the problem of trying to explain the price of houses. The price of a house depends on many characteristics (e.g. number of bedrooms, number of bathrooms, location of house, size of lot, etc.). As in the above example, many variables must be included in a model seeking to explain why some houses are more expensive than others.

These two examples are not unusual. Most problems in applied statistics (e.g. economics, biology, medicine, etc) are of a similar level of complexity. Unfortunately, the basic tool you have encountered so far-simple correlation analysis-cannot handle such complexity. For these more complex cases – that is, those involving more than two variables - regression is the tool to use.

11.1.1 Regression as a Best Fitting Line

As a way of understanding regression, let us begin with just two variables (Y and X). We refer to this case as simple regression. Beginning with simple regression makes sense since graphical intuition can be developed in a straightforward manner and the relationship between regression and correlation can be illustrated quite easily.

To start with, we assume that a linear relationship exists between Y and X . As an example, you might consider Y to be the house price variable and X to be the lot size variable from data set HPRICE.XLS. Remember that this data set contained the sales price of 546 houses in Windsor, Canada along with several characteristics for each house. It is sensible to assume that the size of the lot affects the price at which a house sells.

We can express the linear relationship between Y and X mathematically as:

$$Y = \alpha + \beta X \quad (11.1)$$

where α is the intercept of the line and β is the slope. This equation is referred to as the **regression line**. If in actuality we knew what α and β were, then we would know what the relationship between Y and X was. In practice, of course, we do not have this information. Furthermore, even if our regression model, which posits a linear relationship between Y and X , were true, in the real world we would never find that our data points lie precisely on a straight line. Factors such as measurement error mean that individual data points might lie close to but not exactly on a straight line.

Example 11.1

For instance, suppose the price of a house (Y) depends on the lot size (X) in the following manner:

$$Y = 34,000 + 7X$$

(i.e. $\alpha = 34,000$ and $\beta = 7$). If X were 5,000 square feet, this model says the price of the house should be $Y = 34,000 + 7 \times 5,000 = \$69,000$. But, of course, not every house with a lot size of 5,000 square feet will have a sales price of precisely \$69,000. No doubt in this case, the regression model is missing some important variables (e.g. number of bedrooms) that may affect the price of a house. Furthermore, the price of some houses might be higher than they should be (e.g. if they were bought by irrationally exuberant buyers). Alternatively, some houses may sell for less than their true worth (e.g. if the sellers have to relocate to a different city and must sell their houses quickly). For all these reasons, even if $Y = 34,000 + 7X$ is an accurate description of a straight line relationship between Y and X , it will not be the case that every data point lies exactly on the line.

Our house price example illustrates a truth about regression modeling: **the linear regression model will always be only an approximation of the true relationship**. The truth may differ in many ways from the approximation implicit in the linear regression model.

In economics, the most probable source of error is due to missing variables, usually because we cannot observe them. In our previous example, house prices reflect many variables for which we can easily collect data (e.g. number of bedrooms, number of bathrooms, etc.). But they will also depend on many other factors for which it is difficult if not impossible to collect data (e.g. the number of loud parties held by neighbors, the degree to which the owners have kept the property well-maintained, the quality of the interior decoration of the house, etc.). The omission of these variables from the regression model will mean that the model makes an error. We call all such errors e . The regression model can now be written as:

$$Y = \alpha + \beta X + e \quad (11.2)$$

In the regression model, Y is referred to as the **dependent variable**, X the **explanatory variable**, and α and β , **coefficients**. It is common to implicitly assume that the explanatory variable “causes” Y , and the coefficient b measures the influence of X on Y . In light of the comments made in the previous Unit 3 about how correlation does not necessarily imply causality,

you may want to question the assumption that the explanatory variable causes the dependent variable. There are three responses that can be made to this statement.

First, note that we talk about the regression model. A model specifies how different variables interact. For instance, models of land use posit that population pressures cause rural farmers to expand their lands by cutting down forests, thus causing deforestation. Such models have the causality “built-in” and the purpose of a regression involving $Y = \text{deforestation}$ and $X = \text{population density}$ is to measure the magnitude of the effect of population pressures only (i.e. the causality assumption may be reasonable and we do not mind assuming it).

Secondly, we can treat the regression purely as a technique for generalizing correlation and interpret the numbers that the regression model produces purely as reflecting the association between variables. In other words, we can drop the causality assumption if we wish.

Thirdly, we can acknowledge that the implicit assumption of causality can be a problem and develop new methods.

11.1.2 Errors and Residuals

In light of the error, e , and the fact that we do not know what α and β are, the first problem in regression analysis is how we can figure approximately, or estimate, what α and β are. It is standard practice to refer to the estimates of α and β as $\hat{\alpha}$ and $\hat{\beta}$ (i.e. $\hat{\alpha}$ and $\hat{\beta}$ are actual numbers that the computer calculates, for instance, $\hat{\alpha} = 34,136$ and $\hat{\beta} = 6.599$, which are estimates of the unknown true values $\alpha = 34,000$ and $\beta = 7$). In practice, the way we find estimates is by drawing a line through the points on an XY -plot which fits best. Hence, we must define what we mean by “best fitting line”.

Before we do this, it is useful to make a distinction between errors and residuals. The error is defined as the distance between a particular data point and the true regression line. Mathematically, we can rearrange the regression model to write $e_i = Y_i - \alpha - \beta X_i$. This is the error for the i th observation. However, if we replace α and β by their estimates $\hat{\alpha}$ and $\hat{\beta}$, we get a straight line which is generally a little different from the true regression line. The deviations from this estimated regression line are called **residuals**. We will use the notation “ u ” when we refer to residuals. That is, the residuals are given by $u_i = Y_i - \hat{\alpha} - \hat{\beta} X_i$. If you find the distinction between errors and residuals confusing, you can probably ignore it in the rest of this book and assume errors and residuals are the same thing. However, if you plan on further study of econometrics, this distinction becomes crucial.

If we return to some basic geometry, note that we can draw one (and only one) straight line connecting any two distinct points. Thus, in the case of two points, there is no doubt about what the best fitting line through an XY -plot is. However, typically we have many points—for instance, our deforestation/population density example has 70 different countries and the XY -plots 70 points—and there is ambiguity about what is the “best fitting line”.

Figure 11.1 plots three data points (A, B and C) on an XY graph. Clearly, there is no straight line that passes through all three points. The line I have drawn does not pass through any of them; each point, in other words, is a little bit off the line. To put it another way: the line drawn implies residuals that are labeled u_1 , u_2 and u_3 . The residuals are the vertical difference between a data point and the line. A good fitting line will have small residuals.

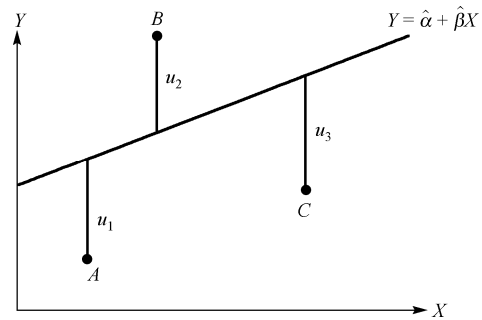


Figure 11.1 Best fitting line for three data points

The usual way of measuring the size of the residuals is by means of the **sum of squared residuals (SSR)**, which is given by: for $i = 1, \dots, N$ data points. We want to find the best fitting line which minimizes the sum of squared residuals. For this reason, estimates found in this way are called **least squares estimates** (or **ordinary least squares-OLS**-to distinguish them from more complicated estimators which we will not discuss until the last unit of this book).

In practice, software packages such as Excel can automatically find values for $\hat{\alpha}$ and $\hat{\beta}$ which will minimize the sum of squared residuals. The exact formulae for $\hat{\alpha}$ and $\hat{\beta}$ can be derived using simple calculus, but we will not derive them here.

Example 11.2

The regression of deforestation on population density. Consider again the data set FOREST.XLS, which contains data on population density and deforestation for 70 tropical countries. It makes sense to assume that population density influences deforestation rather than the other way around. Thus we choose deforestation as the dependent variable (i.e. Y = deforestation) and population as the explanatory variable (i.e. X = population density). Using Excel (Tools/Data Analysis/ Regression) we obtain $\hat{\alpha} = 0.60$ and $\hat{\beta} = 0.000842$. To provide some more jargon, note that when we estimate a regression model it is common to say that “we run a regression of Y on X ”.

Note also that it is actually very easy to calculate these numbers in most statistical software packages. Appropriately, we will turn instead to the more important issue: how we interpret these numbers.

Example 11.3

The effect of advertising on sales The file ADVERT.XLS contains data on annual sales and advertising expenditures (both measured in millions of dollars) for 84 companies in the US. A company executive might be interested in trying to quantify the effect of advertising on sales. This suggests running a regression with dependent variable Y = sales and explanatory variable X = advertising expenditures. Doing so, we obtain the value $\hat{\alpha} = 502.02$ and $\hat{\beta} = 0.218$, which is indicative of a positive relationship between advertising and sales.

New Words and Expressions

- complain [kəm'pleɪn] *vt.* 诉说, 申诉, 控告[后面常跟从句]
layoff ['leɪɔ:f] *n.* 停工, 停止活动; 临时解雇; 操作停止; 失业期
at stake [æt steɪk] *adv.* 危如累卵, 危险
lot [lɒt] *n.* 签, 阄; 份额; 许多; (出售或拍卖的) 一件(货品或商品)
exuberant [ɪg'zju:bərənt] *adj.* 生气勃勃的; (活力) 充沛的; 茂盛的, 繁茂的
rural ['ruərəl] *adj.* 乡下的, 农村的; 田园的; 地方的; 农业的
forest ['fɒrɪst] *n.* 森林; 丛林; (森林似的) 一丛; 一片
deforestation [ˌdi:fɒr'steɪʃn] *n.* 采伐森林, 森林开伐
residual [rɪ'zɪdʒuəl] *n.* 剩余; 残渣; [统计]残差 *adj.* 残余的; 残留的
confuse [kən'fju:z] *vt.* 使困窘; 使混乱; 使困惑; 使更难以理解
tropical ['trɒpɪkl] *adj.* 热带的; 炎热的; 热情的
jargon ['dʒɑ:gən] *n.* 行话; 行业术语; 黑话
expenditure [ɪk'spendɪtʃə(r)] *n.* 花费, 支出; 费用, 经费; (尤指金钱的) 支出额

Technical Terms

- fitting line 拟合直线
best fitting line 最佳拟合直线
least squares 最小二乘法, 最小二乘平方(法)
least square estimates 最小二乘(平方)估计
ordinary least squares (OLS) 普通最小二乘法, 或普通最小二乘方法

Notes

1. residual analysis 残差分析; residual value 残值; residual variance 残差方差
2. Windsor, 即温莎市, 位于加拿大安大略省, 距离安大略首府多伦多市约 4 小时车程, 而到离隔河相望的北美洲汽车城底特律则只需 5 分钟车程。

11.2 Interpreting OLS Estimates

In the previous example of the relationship between deforestation and population density, we obtained OLS estimates for the intercept and slope of the regression line. The question now arises: how should we interpret these estimates? The intercept in the regression model, α , usually has little economic interpretation so we will not discuss it here. However, β is typically quite important. This coefficient is the slope of the best fitting straight line through the XY -plot. In the deforestation/population density example, $\hat{\beta}$ was positive.

Remembering the discussion on how to interpret correlations in the previous unit 3, we note

that since $\hat{\beta}$ is positive X and Y are positively correlated. However, we can go further in interpreting $\hat{\beta}$ if we differentiate the regression model and obtain:

$$\frac{dY}{dX} = \beta \quad (11.3)$$

Even if you do not know calculus, the verbal intuition of the previous expression is not hard to provide. Derivatives measure how much Y changes when X is changed by a small (marginal) amount. Hence, β can be interpreted as the marginal effect of X on Y and is a measure of how much X influences Y .

To be more precise, we can interpret β as a measure of how much Y tends to change when X is changed by one unit. The definition of “unit” in the previous sentence depends on the particular data set being studied and is best illustrated through examples. Before doing this, it should be stressed that regressions measure tendencies in the data (note the use of the word “tends” in the explanation of β above). It is not necessarily the case that every observation (e.g. country or house) fits the general pattern established by the other observations. In this case we called such unusual observations outliers and argued that, in some cases, examining outliers could be quite informative.

In the case of regression, outliers are those with residuals that stand out as being unusually large. Hence, examining the residuals from a regression is a common practice. (In Excel you can examine the residuals by clicking on the box labeled “Residuals” in the regression menu.)

Example11.4 (Example11.2 continued)

In the deforestation/population density example we obtained $\hat{\beta} = 0.000842$. This is a measure of how much deforestation tends to change when population density changes by a small amount. Since population density is measured in terms of the number of people per 1,000 hectares and deforestation as the percentage forest loss per year, this figure implies that if we add one more person per 1,000 hectares (i.e. a change of one unit in the explanatory variable) deforestation will tend to increase by 0.000842%.

Alternatively, we could present this information as follows. The population density varies quite a bit across countries: from below 100 people to over 2,500 people per 1,000 hectares. Hence it is not surprising that a change of one person per hectare will have little effect on deforestation. We could multiply everything by 100 and say that “increasing population density by 100 people per thousand hectares will tend to increase deforestation by 0.0842%”. Even the latter number may seem insignificant, but note that an increase of annual deforestation rates by 0.0842% per year will result in a country losing an extra 5% of its forest over 50 years. In the long run and over a large area-the spatial and time scales in which environmental economists are accustomed to thinking-this degree of forest loss can be substantial.

Example11.5 (Example11.3 continued)

Both advertising and sales are measured in millions of dollars and we found $\hat{\beta} = 0.218$. Following the same line of reasoning above, we can say that a one million dollar increase in

advertising tends to be associated with a \$218,000 increase in sales (i.e. $1,000,000 \times 0.218 = 218,000$). This result would seem to indicate that spending on advertising is rather counter productive since an extra \$1,000,000 spent on advertising would only translate into an extra \$218,000 in sales.

Does this mean that the company executive running this regression should decide to reduce advertising expenditures? Possibly, but not necessarily. The reason for this uncertainty relates to the issue of causality and the question of how correlation or regression results can be interpreted (see Unit 3). That is, if the regression truly is a causal one (i.e. it is the case that advertising has a direct influence on sales), then we can interpret the \$218,000 figure as indicative of what the effect of a change in advertising will be.

However, if it is not causal, then it is risky to use the regression result to provide strategic advice to a company. Indeed, it is possible that larger companies tend to have egomaniacs as bosses and egomaniacs enjoy seeing their companies advertised. If this (possibly implausible) story is true then we would expect to see larger companies advertising more-exactly what our regression has found. Such an interpretation would imply that it is possible that advertising is not directly influencing sales. The apparent positive relationship between advertising and sales from the regression analysis may be due solely to the behavior of the bosses of large companies.

Deciding whether it is reasonable to assume that a regression model captures a causal relationship in which one variable directly influences another is very difficult, and it is hard to offer any general rules on the subject.

New Words and Expressions

outlier ['aʊt,laɪə] *n.* 露宿者；局外人；[数]离群值；异常值

hectare ['hekteə(r)] *n.* 公顷。复数 hectares

extra ['ekstrə] *adj.* 额外的，补充的，附加的；特大的，特别的

n. 附加物，额外的事物；临时演员；上等产品

egomaniacs [i:gəʊ'meɪniæk] *n.* 极端利己主义者，极端自我主义者

insignificant [ˌɪnsɪg'nɪfɪkənt] *adj.* 不重要的；微小的；毫无意义的；不足道

Technical Terms

statistically insignificant 统计上不显著

insignificant difference 区别不明显

insignificant digit 无效数字 (significant digit 有效数字)

11.3 Fitted Values and R^2 : Measuring the Fit of a Regression Model

In the preceding discussion we learned how to calculate and interpret regression coefficients, $\hat{\alpha}$ and $\hat{\beta}$. Furthermore, we explained that regression finds the “best fitting” line in the sense that

it minimizes the SSR. However, it is possible that the “best” fit is not a very good fit at all. Appropriately, it is desirable to have some measure of fit (or a measure of how good the best fitting line is). The most common measure of fit is referred to as the R^2 . It relates closely to the correlation between Y and X . In fact, for the simple regression model, it is the correlation squared. This provides the formal statistical link between regression and correlation. However, the previous discussion should make the informal links between correlation and regression clear. Both are interested in quantifying the degree of association between different variables and both can be interpreted in terms of fitting lines through XY -plots.

To derive and explain R^2 , we will begin with some background material. We start by clarifying the notion of a fitted value. Remember that regression fits a straight line through an XY -plot, but does not pass precisely through each point on the plot (i.e. an error is made). In the case of our deforestation/population density example, this meant that individual countries did not lie on the regression line. The fitted value for observation i is the value that lies on the regression line corresponding to the X_i value for that particular observation (e.g. house, country). In other words, if you draw a straight vertical line through a particular point in the XY -plot, the intersection of this vertical line and the regression line is the fitted value corresponding to the point you chose.

Alternatively, we can think of the idea of a fitted value in terms of the formula for the regression model:

$$Y_i = \alpha + \beta X_i + e_i \quad (11.4)$$

Remember that adding i subscripts (e.g. Y_i) indicates that we are referring to a particular observation (e.g. the i th country or the i th house). If we ignore the error, we can say that the model’s prediction of Y_i should be equal to $\alpha + \beta X_i$. If we replace α and β by the OLS estimates $\hat{\alpha}$ and $\hat{\beta}$, we obtain a so-called “fitted” or “predicted” value for Y_i :

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} \hat{X}_i \quad (11.5)$$

Note that we are using the value of the explanatory variable and the OLS estimates to predict the dependent variable. By looking at actual (Y_i) versus fitted (\hat{Y}_i) values we can gain a rough impression of the “goodness of fit” of the regression model.

Many software packages allow you to print out the actual and fitted values for each observation. An examination of these values not only gives you a rough measure of how well the regression model fits, they allow you to examine individual observations to determine which ones are close to the regression line and which are not. Since the regression line captures general patterns or tendencies in your data set, you can see which observations conform to the general pattern and which do not.

We have defined the residual made in fitting our best fitting line previously. Another way to express this residual is in terms of the difference between the actual and fitted values of Y . That is:

$$u = Y_i - \hat{Y}_i \quad (11.6)$$

Software packages such as Excel can also plot or list the residuals from a regression model. These

can be examined in turn to give a rough impression of the goodness of fit of the regression model. We emphasize that unusually big residuals are outliers and sometimes these outliers are of interest.

To illustrate the kind of information with which residual analysis can provide us, take a look at your computer output from Problem 11.3 (a). In the Residual Output, observation 39 has a fitted value of 2.93 and a residual of -1.63 . By adding these two figures together (or by looking at the original data), you can see that the actual deforestation rate for this country is 1.3. What do all these numbers imply?

Note that the regression model is predicting a much higher value (2.93) for deforestation than actually occurred (1.3) in this country. This means that this country may be doing much better at protecting its forests than the regression model implies, and, consequently, is making better efforts at forest conservation than are other countries. This kind of information may be important to policymakers in other countries, particularly as this outlier country may provide useful lessons that can be applied to them.

The ideas of a residual and a fitted value are important in developing an informal understanding of how well a regression model fits. However, we still lack a formal numerical measure of fit. At this stage, we can now derive and motivate such a measure: R^2 .

Recall that variance is the measure of dispersion or variability of the data. Here we define a closely related concept, the **total sum of squares** or TSS:

$$\text{TSS} = \sum (Y_i - \bar{Y})^2 \quad (11.7)$$

Note that the formula for the variance of Y is $\text{TSS}/(N-1)$ (see Unit 2). Loosely speaking, the $N-1$ term will cancel out in our final formula for R^2 and, hence, we ignore it. So think of TSS as being a measure of the variability of Y . The regression model seeks to explain the variability in Y through the explanatory variable X . It can be shown that the total variability in Y can be broken into two parts as:

$$\text{TSS} = \text{RSS} + \text{SSR} \quad (11.8)$$

where RSS is the **regression sum of squares**, a measure of the explanation provided by the regression model. RSS is given by:

$$\text{RSS} = \sum (\hat{Y}_i - \bar{Y})^2 \quad (11.9)$$

Remembering that SSR is the **sum of squared residuals** and that a good fitting regression model will make the SSR very small, we can combine the equations above to yield a measure of fit:

$$R^2 = 1 - \frac{\text{SSR}}{\text{TSS}} \quad (11.10)$$

or, equivalently,

$$R^2 = \frac{\text{RSS}}{\text{TSS}} \quad (11.11)$$

Intuitively, the R^2 measures the proportion of the total variance of Y that can be explained by X . Note that TSS, RSS and SSR are all sums of squared numbers and, hence, are all non-negative.

This implies $TSS \geq RSS$ and $TSS \geq SSR$. Using these facts, it can be seen that $0 \leq R^2 \leq 1$.

Further intuition about this measure of fit can be obtained by noting that small values of SSR indicate that the regression model is fitting well. A regression line which fits all the data points perfectly in the XY -plot will have no errors and hence $SSR = 0$ and $R^2 = 1$. Looking at the formula above, you can see that values of R^2 near 1 imply a good fit and that $R^2 = 1$ implies a perfect fit. In sum, high values of R^2 imply a good fit and low values a bad fit.

An alternative source of intuition is provided by the RSS. RSS measures how much of the variation in Y the explanatory variables explain. If RSS is near TSS, then the explanatory variables account for almost all of the variability and the fit will be a good one. Looking at the previous formula you can see that the R^2 is near one in this case.

Example 11.6 (Example 11.3 continued)

The R^2 from the regression of sales on advertising expenditures using data set ADVERT.XLS is 0.09. This relatively small number indicates that variations in advertising expenditures across companies account for only a small proportion of the variation in sales. This finding is probably reasonable, in that you would expect factors other than advertising (e.g. product quality, pricing, etc.) to play a very important role in explaining the sales of a company.

New Words and Expressions

subscript ['sʌbskript] *n.* 下标, 脚注, 下角数码 *adj.* 下标的, 写在下方的, 脚注的

rough [rʌf] *adj.* 粗糙的, 崎岖不平的; 狂暴的, 汹涌的; 未经加工的

n. 粗糙的部分; 艰难, 苦难

impression [ɪm'preʃn] *n.* 印象, 感觉; 影响, 效果; 盖印, 印记

lesson ['lesn] *n.* 教训, 训诫; 功课; 课程, 一堂课 *vt.* 教训, 训斥; 教课; 向.....授课

11.4 Nonlinearity in Regression

So far, we have used the linear regression model and fit a straight line through XY -plots. However, this may not always be appropriate. Consider the XY -plot in Figure 11.2. It looks like the relationship between Y and X is not linear. If we were to fit a straight line through the data, it might give a misleading representation of the relationship between Y and X . In fact, we have artificially generated this data by assuming the relationship between Y and X is of the form:

$$Y_i = 6X_i^2$$

such that the true relationship is quadratic. A cursory glance at the XY -plots can often indicate whether fitting a straight line is appropriate or not.

What should you do if a quadratic relationship rather than a linear relationship exists? The answer is surprisingly simple: rather than regressing Y on X , regress Y on X^2 instead.

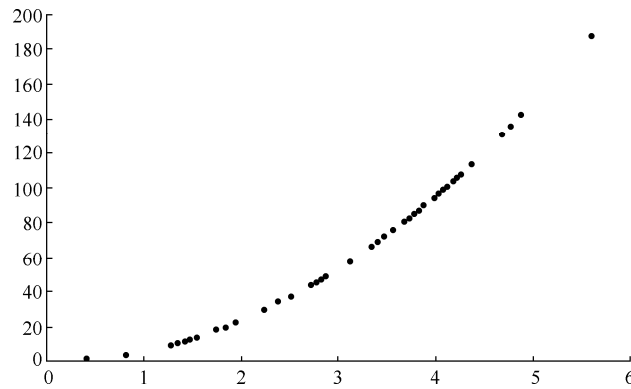


Figure 11.2 A quadratic relationship between X and Y .

Of course, the relationship revealed by the XY -plot may be found to be neither linear nor quadratic. It may appear that Y is related to $\ln(X)$ or $1/X$ or X^3 or any other transformation of X . However, the same general strategy holds: transform the X variable as appropriate and then run a regression of Y on the transformed variable. You can even transform Y if it seems appropriate.

A very common transformation, of both the dependent and explanatory variables, is the logarithmic transformation. Even if you are not familiar with logarithms, they are easy to work with in any spreadsheet or econometric software package, including Excel. Often economists work with natural logarithms, for which the symbol is \ln . In this book, we will always use natural logarithms and simply refer to them as “logs” for short. It is common to say that: “we took the log of variable X ” or that “we worked with $\log X$ ”. The mathematical notation is $\ln(X)$.

Why is it common to use $\ln(Y)$ as the dependent variable and $\ln(X)$ as the explanatory variable? First, the expressions will often allow us to interpret results quite easily.

Second, data transformed in this way often does appear to satisfy the linearity assumption of the regression model.

To fully understand the first point, we need some background in calculus, which is beyond the scope of this book. Fortunately, the intuition can be stated verbally. In the following regression:

$$\ln(Y) = \alpha + \beta \ln(X) + e$$

β can be interpreted as an *elasticity*. Recall that, in the basic regression without logs, we said that “ Y tends to change by β **units** for a one unit change in X ”. In the regression containing both logged dependent and explanatory variables, we can now say that “ Y tends to change by β **percent** for a one **percent** change in X ”. That is, instead of having to worry about units of measurements, regression results using logged variables are always interpreted as *elasticities*. Logs are convenient for other reasons too.

For instance, as discussed in Unit 2, when we have time series data, the percentage change in a variable is approximately $100 \times [\ln(Y_t) - \ln(Y_{t-1})]$. This transformation will turn out to be useful.

The second justification for the log transformation is purely practical: With many data sets, if you take the logs of dependent and explanatory variables and make an XY -plot the resulting relationship will look linear. This is illustrated in Figures 11.3 and 11.4. Figure 11.3 is an XY -plot

of two data series, Y and X , neither of which has been transformed in any way. Figure 11.4 is an XY -plot of $\ln(X)$ and $\ln(Y)$.

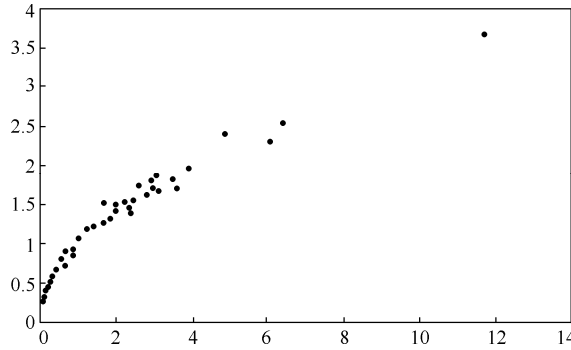


Figure 11.3 X and Y need to be logged

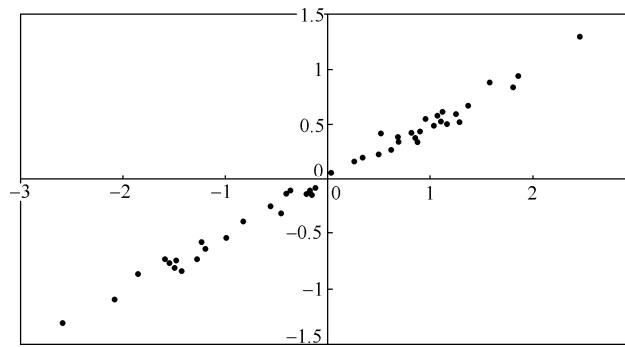


Figure 11.4 $\ln(X)$ versus $\ln(Y)$

Note that the points in the first figure do not seem to lie along a straight line. Rather the relationship is one of a steep-sloped pattern for small values of X , that gradually flattens out as X increases. This is a typical pattern for data which should be logged. Figure 11.4 shows that, once the data is logged, the XY -plot indicates a linear pattern. An OLS regression will fit a straight line with a high degree of accuracy in Figure 11.4. However, fitting an accurate straight line through Figure 11.3 is a very difficult (and probably not the best) thing to do.

On what basis should you log your data (or for that matter take any other transformation)? There is no simple rule that can be given. Examining XY -plots of the data transformed in various ways is often instructive. For instance, begin by looking at a plot of X against Y . This may look roughly linear. If so, just go ahead and run a regression of Y on X . If the plot does not look linear, it may exhibit some other pattern that you recognize (e.g. the quadratic form of Figure 11.2 or the logarithmic form of Figure 11.3). If so, create an XY -plot of suitable transformed variables (e.g. $\ln(Y)$ against $\ln(X)$) and see if it looks linear. Such a strategy will likely work well in a simple regression containing only one explanatory variable.

In these cases, the examination of XY -plots may be quite complicated since there are so many possible XY -plots that could be constructed.

New Words and Expressions

quadratic [kwɒ'drætik] *adj.* 二次的 *n.* 二次方程式

cursory ['kɜ:səri] *adj.* 粗略的, 草率的, 仓促的; 肤皮潦草

elasticity [i:læ'stisəti] *n.* 弹性; 弹力; 灵活性; 伸缩性

logarithm ['lɒgərɪðəm] *n.* 对数

flatten ['flætən] *vt. & vi.* 变平, 使(某物)变平; 打倒, 击倒; 使失去光泽

Technical Terms

logarithmic transformation 对数变换

logged variables 取对数变量

common logarithm 常用对数

natural logarithm 自然对数

Notes

1. steep-sloped 陡峭的斜坡

Summary

1. Simple regression quantifies the effect of an explanatory variable, X , on a dependent variable, Y . Hence, it measures the relationship between two variables.

2. The relationship between Y and X is assumed to take the form, $Y = \alpha + \beta X$, where α is the intercept and β the slope of a straight line. This is called the regression line.

3. The regression line is the best fitting line through an XY graph.

4. No line will ever fit perfectly through all the points in an XY graph. The distance between each point and the line is called a residual.

5. The ordinary least squares (OLS) estimator is the one which minimizes the sum of squared residuals.

6. OLS provides estimates of α and β which are labeled $\hat{\alpha}$ and $\hat{\beta}$.

7. Regression coefficients should be interpreted as marginal effects (i.e. as measures of the effect on Y of a small change in X).

8. R^2 is a measure of how well the regression line fits through the XY graph.

9. OLS estimates and the R^2 are calculated in computer software packages such as Excel.

10. Regression lines do not have to be linear. To carry out nonlinear regression, merely replace Y and/or X in the regression model by a suitable nonlinear transformation (e.g. $\ln(Y)$ or X^2).

Reading English Materials

Passage: Mathematical details of OLS

The OLS estimator defines the best fitting line through the points on an XY -plot. Mathematically, we are interested in choosing $\hat{\alpha}$ and $\hat{\beta}$ so as to minimize the sum of squared residuals. The SSR can be written as:

Optional exercise

Take first and second derivatives with respect to $\hat{\alpha}$ and $\hat{\beta}$ of the above expression for SSR. Use these to find values of $\hat{\alpha}$ and $\hat{\beta}$ that minimize SSR. Verify that the solution you have found does indeed minimize (rather than maximize) SSR.

If you have done the previous exercise correctly, you should have obtained the following:

$$\hat{\beta} = \frac{\sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

and

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

where \bar{Y} and \bar{X} are the means of Y and X . These are the OLS estimators for α and β . Note that there are several equivalent ways of writing the formula for $\hat{\beta}$. If you consult other textbooks you will find alternative expressions for the OLS estimator.

These equations can be used to demonstrate the consequences of taking deviations from means. By way of explanation, note that we have assumed above that the dependent and explanatory variables, X and Y , are based on the raw data.

However, in some cases researchers do not work with just X and Y , but rather with X and Y minus their respective means:

$$y_i = Y_i - \bar{Y}$$

and

$$x_i = X_i - \bar{X}$$

Consider using OLS to estimate the regression:

$$y = a + bX + e$$

where we have used the symbols a and b to distinguish them from the coefficients α and β in the regression involving Y and X .

It turns out that the relationship between OLS estimates from the original regression and the one where deviations from means have been taken is a simple one. The OLS estimate of β is always exactly the same as $\hat{\beta}$ and the OLS estimate of a is always zero. In other words, taking

deviations from means simplifies the regression model by getting rid of the intercept (i.e. there is no point in including an intercept since its coefficient is always zero). This simplification does not have any effect on the slope coefficient in the regression model. It is unchanged by taking deviations from means and still has the same interpretation as a marginal effect.

It is not too hard to prove the statements in the previous paragraph and, if you are mathematically inclined, you might be interested in doing so. As a hint, note that the means of y and x are zero.

In this case, if you take deviations from means of the dependent and all of the explanatory variables, you obtain the same result. That is, the intercept disappears from the regression, but all other coefficient estimates are unaffected.

Problems

11.1 The Excel data set FOREST.XLS contains data on Y = deforestation, X = population density, W = change in cropland(农田) and Z = change in pasture land(牧场).

- Run a regression of Y on X and interpret the results.
- Run a regression of Y on W and one of Y on Z and interpret the results.
- Create a new variable, V , by dividing X by 100. What are the units in terms of which V is measured?
- Run a regression of Y on V . Compare your results to those for (a). How do you interpret your coefficient estimate of (b)? How does $\hat{\alpha}$ differ between (a) and (d)?

11.2 Using the data in FOREST.XLS (see Problem 11.1), run a regression of Y on X using Excel with the box clicked on labeled “Line Fit Plot” in the regression menu. Graphically and numerically compare the actual to the fitted values (i.e. look at the columns labeled “Residual Output” and the accompanying display chart).

11.3 a. Using the data in FOREST.XLS (see Problem 11.1) run a regression of Y on X using Excel with the boxes labeled “Residuals” and “Residual Plots” in the regression menu clicked on. How would you interpret the residuals? Are there any outliers?

- Repeat question part a for the other variables, W and Z in this data set.

11.4 a. Using the data in FOREST.XLS (see Problem 11.1), run a regression of Y on X using Excel. What is the R^2 ?

- Calculate the correlation between Y and X .
- Discuss the relationship between your answers in part a and part b.
- Redo part a for various regressions involving the variables W , X , Y and Z in the data set. Comment on the fit of each of these regressions.

11.5 Using the data in FOREST.XLS examine different XY -plots involving the variables X , Y , W and Z (see Problem 11.1 for a definition of these variables). Does there seem to be a nonlinear relationship between any pair of variables? Repeat the exercise using the data in the advertising example (ADVERT.XLS).

11.6 Data set EX46.XLS contains two variables, labeled Y and X .

a. Make an XY -plot of these two variables. Does the relationship between Y and X appear to be linear?

b. Calculate the square root of variable X . Note the Excel symbol for square root is SQRT.

c. Make an XY -plot of the square root of X against Y . Does this relationship appear to be linear?

11.7 Use the data in the example related to costs of production in the electric utility industry (ELECTRIC.XLS), where Y = cost of production(生产成本) and X = output(产出).

a. Run a regression of Y on X .

b. Take log transformations of both variables.

c. Run a regression of $\ln(Y)$ on $\ln(X)$ and interpret your results verbally.

Part III Statistical Methods and Data Science

I keep saying that the sexy job in the next 10 years will be statisticians.

——Hal Varian, Chief Economist at Google

Data Scientist: The Sexiest Job of the 21st Century

——Thomas H. Davenport and D.J. Patil

Thomas H. Davenport is a distinguished professor at Babson College, a research fellow at the MIT Center for Digital Business, and a senior adviser to Deloitte Analytics. He is at work on a book about automation in knowledge work.

D.J. Patil is the data scientist in residence at Greylock Partners, was formerly the head of data products at LinkedIn, and is the author of *Data Jujitsu: The Art of Turning Data into Product* (O'Reilly Media, 2012).



Unit 12

Statistics and Data Science



12.1 Statistics and Data Science (I)



12.2 Statistics and Data Science (II)



12.3 Statistical Thinking



12.4 Distinguishing Analytics, Business Intelligence, Data Science

12.1 Statistics and Data Science (I)

12.1.1 What is Data Science

Recently, there has been much hand-wringing about the role of statistics in data science. Statistics is a crucial part of data science, but at the same time, most statistics departments are at grave risk of becoming irrelevant. Statistics is flourishing; by-and-large academic statistics continues to focus on problems that are not relevant to most data analyses. Data science isn't just statistics, and highlight important parts of data science that are typically considered to be out of bounds for statistics research.

Data science is an emerging interdisciplinary field that combines elements of mathematics, statistics, computer science, and knowledge in a particular application domain for the purpose of extracting meaningful information from the increasingly sophisticated array of data available in many settings. These data tend to be nontraditional, in the sense that they are often live, large, complex, and/or messy.

Data science is all the rage in the media. Most people would agree that it has three pillars: computer science, statistics/mathematics and domain knowledge. Some believe data science is the intersection of the three as shown in Figure 12.1 while others think it is the union.

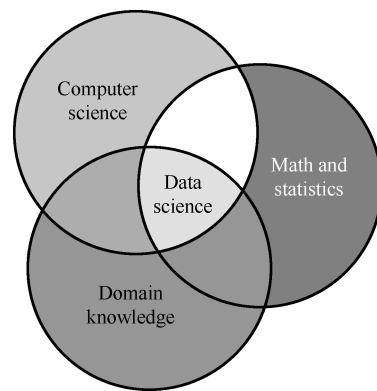


Figure 12.1 Data science: three pillars
(From <http://bulletin.imstat.org/2014/10/ims-presidential-address-let-us-own-data-science/>)

There are three main steps in a data science project: you collect data (and questions), analyze it (using visualization and models), then communicate the results. It's rare to walk this process in one direction: often your analysis will reveal that you need new or different data, or when presenting results you'll discover a flaw in your model.

12.1.2 Statistics and Data Science

A first course in statistics at the undergraduate level typically introduces students to a variety of techniques to analyze small, neat, and clean datasets. However, whether they pursue more formal training in statistics or not, many of these students will end up working with data that are considerably more complex, and will need facility with statistical computing techniques. More importantly, these students require a framework for thinking structurally about data.

However, the data that many of our current students will be asked to analyze—especially if they go into government or industry—will not be so neat and tidy. Indeed, these data are not likely to come from an experiment—they are much more likely to be observational. Second, they will not

likely come in a two-dimensional row-and-column format—they might be stored in a database, or a structured text document (e.g., XML), or come from more than one source with no obvious connecting identifier, or worse, have no structure at all (e.g., data scraped from the web). These data might not exist at a fixed moment in time, but rather be part of a live stream (e.g., Twitter). These data might not even be numerical, but rather consist of text, images, or video. Finally, these data may consist of so many observations that many traditional inferential techniques might not make sense to use, or even be computationally feasible.

But while this data onslaught has strengthened interest in statistics, it has also brought challenges. Modern data streams are importantly different than the data with which many statisticians, and in turn many statistics students, are accustomed to working. For example, the typical dataset a student encounters in an introductory statistics course consists of a several dozen rows and three or four columns of non-collinear variables, collected from a simple random sample or a randomized trial. These are data that are likely to meet the conditions necessary for statistical inference in a multiple regression model.

Statistics has a lot to say about collecting data: survey sampling and design of experiments are well established fields backed by decades of research. Statisticians, however, have little to say about collecting and refining questions. Good questions are crucial for good analysis, but there is little research in statistics about how to solicit and polish good questions, and it's a skill rarely taught in core PhD curricula.

Once the data has been collected, it needs to be tidied (or normalized) into a form that's amenable for analysis. Organizing data into the right 'shape' is essential for fluent data analysis: if it's in the wrong shape you'll spend the majority of your time fighting your tools, not questioning the data. We have worked on this problem for quite some time (culminating in the tidy data framework) but we are aware of little similar work by statisticians.

Any real data analysis involves data manipulation (sometimes called wrangling or munging), visualization and modelling. Visualization and modelling are complementary. Visualizations surprise you, and can help refine vague questions. However, visualizations rely on human interpretation, so the ability to scale is fundamentally constrained. Models scale much better, and it's usually possible to throw more computing at the problem. But models are constrained by their assumptions: fundamentally a model cannot surprise you. In any real analysis you may use both visualizations and models. But the vast majority of statistics research is on modelling, much less is on visualization, and less still on how to iterate between modelling and visualization to get to a good place.

Just as Wilkinson (2006) brought structure to graphics through "grammar", Wickham (2014) and Wickham and Francois (2015) brought structure to data manipulation through the five "verbs": *select*, *filter*, *mutate*, *arrange*, and *summarise*. These common single-table data manipulation operations are the practical descendents of theoretical work on data structures by computer scientists who developed notions of normal forms, relational algebras, and database management systems.

The end product of an analysis is not a model: it is rhetoric. An analysis is meaningless unless

it convinces someone to take action. In business, this typically means convincing senior management who have little statistical expertise. In science, it typically means convincing reviewers. Communication is not a mainstream thread of statistics research (if you attend the JSM, it's easy to come to the conclusion that some academic statisticians couldn't care less about the communication of results). Communication is a part of some PhD programs, but it tends to focus

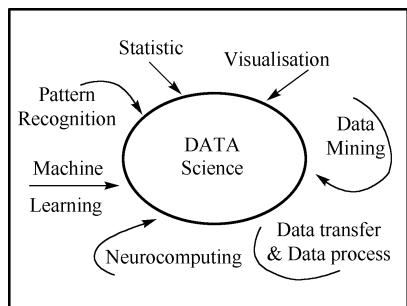


Figure 12.2 Statistics and data science

on professional communication (to other statisticians), not communicating with people who have substantive expertise in other domains.

In business, analyses are often not done just once, but need to be performed again and again as new data come in. These data products need to be robust in both the statistical sense (i.e. to changes in the underlying distributions/assumptions) and in the software engineering sense (i.e. to changes in the underlying technological infrastructure), see Figure 12.2. This is a ripe field for research.

New Words and Expressions

hand-wringing ['hænd'riŋɪŋ] 束手无策

grave [ɡreɪv] *n.* 坟墓, 墓穴; 埋葬.....的地方 *adj.* 重大的, 重要的; 严重的

irrelevant ['ɪrɪləvənt] *adj.* 不相干的; 不恰当; 缺乏时代性的, 落后于潮流的

flourishing ['flɜːʃɪŋ] *adj.* 繁荣的; 欣欣向荣的; 盛行的

interdisciplinary [ˌɪntə'dɪsəplɪnəri] *adj.* 各学科间的; 跨学科

by and large [baɪ ænd lɑːdʒ] 大体上说

pillar ['pɪlə(r)] *n.* (制度、协议等的) 支柱, 核心; 中坚; 栋梁

flaw [flɔː] *n.* 瑕疵, 缺点; 短暂的风暴; 裂缝, 裂纹

tidy ['taɪdi] *adj.* 整洁的, 整齐的; (数量) 相当大的; 相当好的

identifier [aɪ'dentɪfaɪə(r)] *n.* 检验人, 标识符; 鉴别器; 编号

onslaught ['ɒnslɔːt] *n.* 猛攻, 攻击; 突击

accustom [ə'kʌstəm] *vt.* 使习惯

solicit [sə'lɪsɪt] *vt. & vi.* 恳求; 征求; 提起

polish ['pɒlɪʃ] *v.* 擦光; 修改; 润色

amenable [ə'mi:nəbl] *adj.* (对法律等) 负责的; 易控制的; 经得起检验的

manipulation [mə'nɪpjʊ'leɪʃn] *adj.* 操作; 操纵; 控制

mutate [mju:'teɪt] *vt. & vi.* (使某物) 变化; 改变; 突变; 变异

descendent [dɪ'sendənt] *adj.* 下降的; 降落的; 派生的

rhetoric [rɪ'tɒrɪkl] *adj.* 修辞的, 修辞学的; 辞藻华丽的, 虚夸的

thread [θred] *n.* 线; 线索; 线状物

infrastructure ['ɪnfəstrʌktʃə(r)] *n.* 基础设施；基础建设
ripe [raɪp] *adj.* 成熟的；老练的；时机成熟的

Technical Terms

first course in statistics 统计学导论，初级统计学
domain knowledge 特定领域知识

Notes

1. *munge*: (1) verb (used with or without object), munged, munging. Computer Slang. to manipulate (raw data), especially to convert (data) from one format to another, e.g. the munging of HTML content. (2) a noun meaning a comprehensive rewrite of a routine, data structure, or the whole program.

2. *Data munging* or *data wrangling* is loosely the process of manually converting or mapping data from one “raw” form into another format that allows for more convenient consumption of the data with the help of semi-automated tools. This may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses.

Data munging as a process typically follows a set of general steps which begin with extracting the data in a raw form from the data source, “munging” the raw data using algorithms (e.g. sorting) or parsing the data into predefined data structures, and finally depositing the resulting content into a data sink for storage and future use.

3. *data wrangler* is the person performing the wrangling. In the scientific research context, the term often refers to a person responsible for gathering and organizing disparate data sets collected by many different investigators, often as part of a field campaign. In this sense, the term could be credited to Donald Cline during the NASA/NOAA Cold Lands Processes Experiment.

12.2 Statistics and Data Science (II)

12.2.1 Statistics as Part of Data Science

Statistics is a part of data science, not the whole thing. Statistics research focuses on data collection and modelling, and there is little work on developing good questions, thinking about the shape of data, communicating results or building data products.

There are people in statistics doing great work in all these areas, but it's not mainstream statistics. If you're interested in these areas, it's harder to get tenure, harder to get grants, and most of the 'top' statistics journals are unavailable to you.

Attempting to claim that data science is 'just' statistics makes statisticians look out of touch, and belittles the many other contributions outside of statistics, see Figure 12.3.

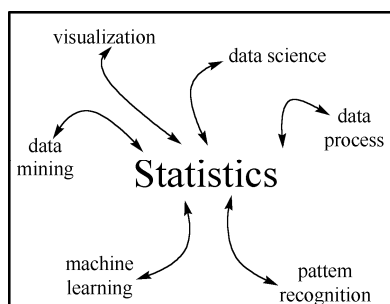


Figure 12.3 One view for the ways that statistics (the science of learning from data) integrates with other key topics. (Source: American Statistical Association.)

12.2.2 The Modern Statistical Analysis Process

Schematic of the modern statistical analysis process. The introductory statistics course (and in many cases, the undergraduate statistics curriculum) emphasizes the central column. In this data science course, we provide instruction into the bubbles to the left and right. See Figure 12.4.

In Figure 12.4, we present a schematic of a modern statistical analysis process, from formulating a question to obtaining an answer. In the introductory statistics course, we teach a streamlined version of this process, wherein challenges with the data, computational methods, and visualization and presentation are typically elided. The entire process informs the material presented in the data science course. The goal is to produce students who have *confidence* and foundational skills — not necessarily expertise — to tackle each step in this modern data analysis cycle, both immediately and in their future careers.

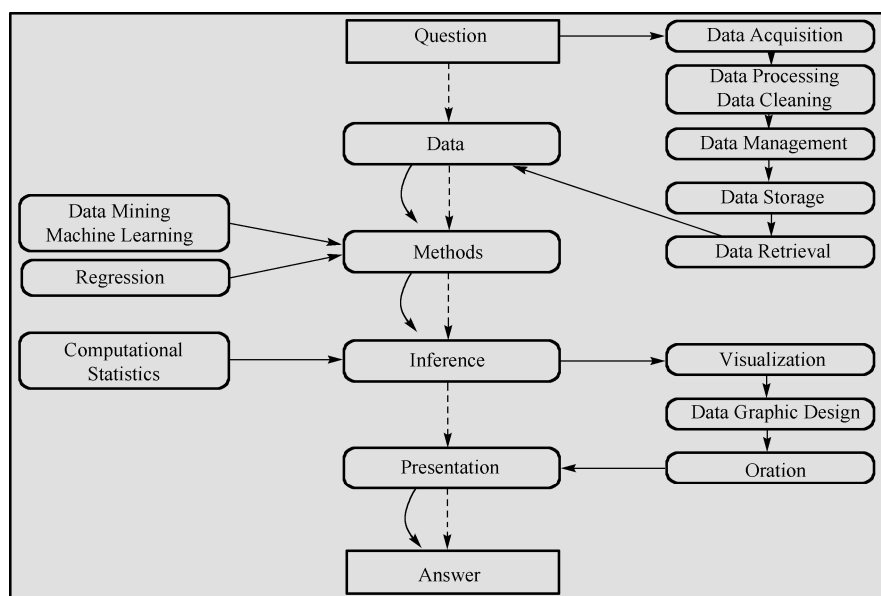


Figure 12.4 Modern statistical analysis process

While the emphasis on computing within the statistics curriculum may be growing, it belongs

to a larger, more gradual evolution in statistics education toward data analysis — with computers — and encourages us to reflect on shifting boundaries between statistics and computer science.

Carnegie Mellon statistics professor Cosma Shalizi considers the differences and similarities between statistics and data science.

If people want to call those who do such jobs “data scientists” rather than “statisticians” because it sounds more dignified, or gets them more money, or makes them easier to hire, then more power to them. If they want to avoid the suggestion that you need a statistics degree to do this work, they have a point but it seems a clumsy way to make it. If, however, the name “statistician” is avoided because that connotes not a powerful discipline which transforms profound ideas about learning from experience into practical tools, but rather, a meaningless conglomeration of rituals better conducted with twenty-sided dice, then we as a profession have failed ourselves and, more importantly, the public, and the blame lies with us. Since what we have to offer is really quite wonderful, we should not let that happen.

Some time during the past couple of years, statistics became data science’s older, more boring sibling that always plays by the rules. There are a lot of statisticians who now call themselves data scientists. I still call myself a statistician.

12.2.3 Statistician and Data Scientist

Big Data are data on a massive scale in terms of volume, intensity, and complexity, and their promise for transforming business, health care, scientific discovery, public policy, and a host of other areas has been proclaimed widely. But, despite the enormous potential for contributions by statisticians, our profession and the ASA (the American Statistical Association) have not been very involved in Big Data activities. We are often missing from Big Data discussions in the media.

There are three reasons for this disconnect. First, the media and public lack a general understanding of what statisticians contribute to society (the issue that motivated the International Year of Statistics). Second, few statisticians are engaged in Big Data projects or have the special skills necessary to handle Big Data challenges.

Third, the statistical community is disconnected from the new (and vaguely defined) community of data scientists, who are completely identified with Big Data in the eyes of the media and policymakers. Data science is frequently described as an amalgam of computer science, mathematics, data visualization, machine learning, distributed data management—and statistics. Data scientists must be innovative modelers and programmers; they also must be exceptional communicators who have a deep understanding of the problem domain and can formulate key questions, uncover novel insights, and use this information to guide high-impact decision making. Other disciplines have been quick to identify themselves with data science and are routinely featured in media accounts. Although statistics is mentioned in passing, statisticians are nearly invisible.

Three Presidents of the ASA wrote on the Statistics and Data Science thing, and eventually faced the same technical asymmetry, see Table 12.1:

Table 12.1 Statistician and Data Scientist

Term	Statistician	Data Scientist
Image	Baseball (Cricket)	HBR Sexiest Job of 21 st Century
Mode	Reactive	Consultative
Works	Solo	In a team
Inputs	Data File, Hypothesis	A Business Problem
Data	Pre-prepared, clean	Distribution, messy, unstructured
Data Size	Kilobytes	Gigabytes
Tools	SAS, SPSS, R, ...	R, Python, awk, Hadoop, Linux, ...
Nouns	Tables	Data Visualizations
Focus	Inference (Why)	Prediction (What)
Output	Report	Data App/ Data Product
Latency	Weeks	Seconds
Stars	G. E. P. Box, Trevor Hastie	Hilary Mason, Nate Silver

Ideally, statistics and statisticians should be the leaders of the Big Data and data science movement. Realistically, we must take a different view. While our discipline is certainly central to any data analysis context, the scope of Big Data and data science goes far beyond our traditional activities.

A quick aside on the “Data Size” row above: while the unstructured or unaggregated data source data that data scientist work with can be in the terabytes range or even large, by the time it’s cleaned and prepared for statistical modeling, a file in the gigabytes range is even more typical.

The Hot Job of the Decade

Hal Varian, the chief economist at Google, is known to have said, “The sexy job in the next 10 years will be statisticians. People think I’m joking, but who would’ve guessed that computer engineers would’ve been the sexy job of the 1990s?”

If “sexy” means having rare qualities that are much in demand, data scientists are already there. They are difficult and expensive to hire and, given the very competitive market for their services, difficult to retain. There simply aren’t a lot of people with their combination of scientific background and computational and analytical skills.

Data scientists today are akin to Wall Street “quants” of the 1980s and 1990s. In those days people with backgrounds in physics and math streamed to investment banks and hedge funds, where they could devise entirely new algorithms and data strategies. Then a variety of universities developed master’s programs in financial engineering, which churned out a second generation of talent that was more accessible to mainstream firms. The pattern was repeated later in the 1990s with search engineers, whose rarefied skills soon came to be taught in computer science programs.

One question raised by this is whether some firms would be wise to wait until that second generation of data scientists emerges, and the candidates are more numerous, less expensive, and easier to vet and assimilate in a business setting. Why not leave the trouble of hunting down and domesticating exotic talent to the big data start-ups and to firms like GE and Walmart, whose aggressive strategies require them to be at the forefront?

The problem with that reasoning is that the advance of big data shows no signs of slowing. If companies sit out this trend's early days for lack of talent, they risk falling behind as competitors and channel partners gain nearly unassailable advantages. Think of big data as an epic wave gathering now, starting to crest. If you want to catch it, you need people who can surf.

New Words and Expressions

- tenure ['tenjə(r)] *n.* 占有（职位，不动产等）；占有期；终身职位
grant [grɑ:nt] *n.* 拨款；补助金；授给物（如财产、授地、专有权、补助、拨款等）
touch [tʌtʃ] *vt.* 触摸；使某物与……轻轻接触；[数]与……相切
out of touch [aut ɔv tʌtʃ] 不联系，不接触，失去联系
belittle [brɪ'lɪtl] *vt.* 轻视；贬低；使显得微小
schematic [ski:'mætɪk] *adj.* 纲要的；示意的；有章法的 *n.* 图表，（尤指）电路原理图
bubble ['bʌbl] *n.* 泡，水泡；冒泡，起泡；泡影，妄想
streamline ['stri:mlaɪn] *vt.* 把……做成流线型；使现代化；使简单化
elide [i'laɪd] *vt.* 忽略，省略（尤指区别）
dignify ['dɪɡnɪfaɪ] *vt.* 使显得威严；使高贵；夸大
clumsy ['klʌmzi] *adj.* 笨拙的；复杂难懂的；（文体等）臃肿的
connote [kə'nəʊt] *vt.* 隐含，暗示，意味着
conglomeration [kən'glɒmə'reɪʃn] *n.* 团块，聚集，混合物
rituals ['rɪtʃʊəlz] *n.* （宗教等的）仪式（ritual 的名词复数）；例行公事，老规矩
blame [bleɪm] *vt.* 指责，责怪；归咎于 *n.* 责备；责任；过失
boring ['bɔ:ɪŋ] *adj.* 无聊的，无趣的；令人厌烦的；单调的，乏味的
proclaim [prə'kleɪm] *vt.* 宣告，公布；表明；赞扬，称颂
disconnect [ˌdɪskə'nekt] *vt.* 切断；断开；分离，使分离；使（电话线路）中断
in the eyes of *adv.* 在……心目中
amalgam [ə'mælgəm] *n.* 汞合金；混合物

Technical Terms

1. schematic map 草图，schematic table 图表题，Data Size 数据量，数据大小
2. ASA (the American Statistical Association) 美国统计协会，美国统计学会
3. International Year of Statistics，国际统计年是指“Statistics 2013”。“Statistics 2013” is a worldwide celebration and recognition of the contributions of statistical science. Through the combined energies of organizations worldwide, Statistics2013 will promote the importance of Statistics to the broader scientific community, business and government data users, the media, policy makers, employers, students, and the general public.

Notes

1. twenty-sided dice : 二十面骰子。



12.3 Statistical Thinking

12.3.1 What is Statistical Thinking

What is statistical thinking? It is well recognized that in order to maintain and improve our competitiveness, we need to continually improve all aspects of our business at an increasing rate. Statistical Thinking provides a common methodology for continuous improvement that is applicable to everything we do. Moreover, Statistical thinking will ensure that we improve efficiently and in a real and lasting manner.

Statistical thinking is the philosophy of learning and action based on the following fundamental principles, see Figure 12.5:

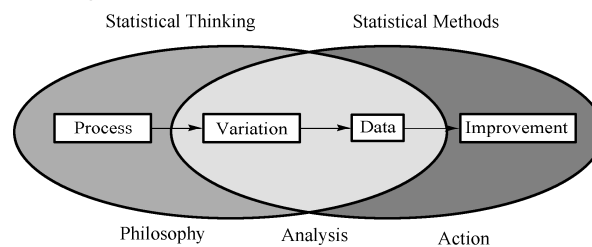


Figure 12.5 Statistical thinking and statistical methods

(i) all work occurs in a system of interconnected processes—a process being a chain of activities that turns inputs into outputs;

(ii) variation, which gives rise to uncertainty, exists in all processes; and

(iii) understanding and reducing variation are keys to success.

All three principles work together to create the power of statistical thinking. The definition highlights several key components: process thinking; understanding and managing uncertainty; and using data whenever possible to guide actions and improve decision-making.

Statistical thinking is a philosophy—a mind-set. It is an overall approach to improvement and therefore more broadly applicable than statistical methods. It is a way of thinking, behaving, working, taking action and interacting with others.

In addition, the process focus of statistical thinking provides the context and the relevancy for broader and more effective use of statistical methods.

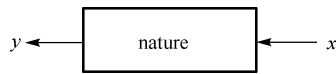
12.3.2 The Two Cultures of Statistical Modeling

There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic

models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems.

Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables x (independent variables) go in one side, and on the other side the response variables y come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;

Information. To extract some information about how nature is associating the response variables to the input variables.

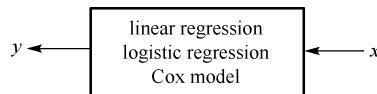
There are two different approaches toward these goals:

(i) The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

$$\text{response variables} = f(\text{predictor variables, random noise, parameters})$$

The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



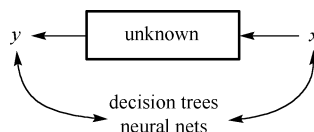
Model validation. Yes-no using goodness-of-fit tests and residual examination.

Estimated culture population. 98% of all statisticians.

(ii) The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(x)$ —an algorithm that operates on x to predict the responses y .

Their black box looks like this:



Model validation. Measured by predictive accuracy.

Estimated culture population. 2% of statisticians, many in other fields.

In the past fifteen years, the growth in algorithmic modeling applications and methodology has been rapid. It has occurred largely outside statistics in a new community — often called *machine learning* — that is mostly young computer scientists.

Under other names, algorithmic modeling has been used by industrial statisticians for decades. See, for instance, the delightful book “*Fitting Equations to Data*” (Daniel and Wood, 1971). It has been used by psychometricians and social scientists. Reading a preprint of Gifi’s book (1990) many years ago uncovered a kindred spirit. It has made small inroads into the analysis of medical data starting with Richard Olshen’s work in the early 1980s. For further work, see Zhang and Singer (1999). Jerome Friedman and Grace Wahba have done pioneering work on the development of algorithmic methods.

But the list of statisticians in the algorithmic modeling business is short, and applications to data are seldom seen in the journals. The development of algorithmic methods was taken up by a community outside statistics.

12.3.3 A New Research Community

In the mid-1980s two powerful new algorithms for fitting data became available: neural nets and decision trees. A new research community using these tools sprang up. Their goal was predictive accuracy. The community consisted of young computer scientists, physicists and engineers plus a few aging statisticians. They began using the new tools in working on complex prediction problems where it was obvious that data models were not applicable: speech recognition, image recognition, nonlinear time series prediction, handwriting recognition, prediction in financial markets.

Their interests range over many fields that were once considered happy hunting grounds for statisticians and have turned out thousands of interesting research papers related to applications and methodology.

A large majority of the papers analyze real data. The criterion for any model is what is the predictive accuracy. An idea of the range of research of this group can be got by looking at the *Proceedings of the Neural Information Processing Systems Conference* (their main yearly meeting) or at the *Machine Learning Journal*.

Theory in Algorithmic Modeling

Data models are rarely used in this community. The approach is that nature produces data in a black box whose insides are complex, mysterious, and, at least, partly unknowable. What is observed is a set of \mathbf{x} ’s that go in and a subsequent set of \mathbf{y} ’s that come out. The problem is to find an algorithm $f(\mathbf{x})$ such that for future x in a test set, $f(\mathbf{x})$ will be a good predictor of \mathbf{y} .

The theory in this field shifts focus from data models to the properties of algorithms. It characterizes their “strength” as predictors, convergence if they are iterative, and what gives them good predictive accuracy. The one assumption made in the theory is that the data is drawn i.i.d. from an unknown multivariate distribution.

There is isolated work in statistics where the focus is on the theory of the algorithms. Grace Wahba's research on smoothing spline algorithms and their applications to data (using cross-validation) is built on theory involving reproducing kernels in Hilbert Space (1990). The final chapter of the CART book (Breiman et al., *Classification and Regression Trees*.1984) contains a proof of the asymptotic convergence of the CART algorithm to the Bayes risk by letting the trees grow as the sample size increases. There are others, but the relative frequency is small.

Theory resulted in a major advance in machine learning. Vladimir Vapnik constructed informative bounds on the generalization error (infinite test set error) of classification algorithms which depend on the "capacity" of the algorithm. These theoretical bounds led to support vector machines (see Vapnik's book, 1995, 1998) which have proved to be more accurate predictors in classification and regression than neural nets, and are the subject of heated current research.

New Words and Expressions

interconnect [ˌɪntəkəˈnekt] *vt.* 使互相连接；使互相联系 *vi.* 互相连接，互相联系

interconnected system 互联系统

philosophy [fəˈlɒsəfi] *n.* 哲学；哲学体系，哲学思想；生活信条；哲理

relevancy [ˈreləvənsi:] *n.* 关联；关联事物；切题的话

culture [ˈkʌltʃə(r)] *n.* 文化，文明（如艺术、哲学）；休养，教养，精神文明

delightful [dɪˈlaɪtfl] *adj.* 令人非常高兴的，讨人喜欢的；令人愉快的

inroad [ˈɪnrəʊd] *n.* 进展；侵袭

spring [sprɪŋ] *vt.* 突然跳出；跳过；使开裂 *spring up* 突然出现，并发；萌芽，长出

psychometrician *n.* 心理测量医生，心理测量专家

smooth [smuːð] *v.* 使平整；使平坦；使平滑；使光滑

ing [ˈsmuːðɪŋ] *v.* (使)光滑，(使)平坦，(使)顺利 (smooth 的现在分词)；缓和；使平和

Technical Terms

1. neural net 神经网络的

2. i.i.d. 是 independent and identically distributed 的缩写，即独立同分布的

3. cross-validation 交叉验证；交叉检验

4. Hilbert Space 希尔伯特空间

Notes

1. handwriting recognition 手写体识别，手写辨识

2. hunting grounds 狩猎场，猎场

12.4 Distinguishing Analytics, Business Intelligence, Data Science

Continuing developments in the fields of Business Intelligence, Analytics, and Data Science are making it increasingly necessary for organizations to become cognizant of the distinctions between these terms, as they relate to the value they can produce for the enterprise. The latest developments to influence these various aspects of the enterprise include:

- ◆ The Simplification of BI: BI vendors are continuing to develop a multitude of tools and technologies that reduces the complexity of BI and its latency while empowering the business user.
- ◆ The Expansion of Analytics: Analytics is progressing into more and more applications and has developed to the point in which it can actually prescribe appropriate and inappropriate actions for specific industries and business units.
- ◆ The Mutability of Data Scientists: It has become increasingly apparent that Data Scientists must exhibit the skills necessary to convert scientific insight regarding data into uses and boons for the business and upper level management for this field to continue to thrive and prove itself.

12.4.1 Analytics

Analytics is probably the single most important aspect of these three frequently confused terms, for the simple fact that both BI and Data Science utilize (and in many cases rely upon) analytics.

Gartner's definition of analytics states that "Analytics has emerged as a catch-all term for a variety of different business intelligence (BI) – and application-related initiatives...Increasingly, 'analytics' is used to describe statistical and mathematical data analysis that clusters, segments, scores and predicts what scenarios are most likely to happen."

At the root of the definition of the term is the fact that analytics hinge upon algorithms to statistically determine relationships between data that can yield insight. The key difference between analytics and BI is that the former has predictive capabilities, whereas the latter has traditionally been based on providing analysis of historical data.

The capability of analytics to determine the likelihood of future events is largely possible through tools such as online analytical processing (OLAP), data mining, forecasting, and data modeling. The process involves analyzing current and historical data patterns to determine future ones, while prescriptive capabilities can analyze future scenarios and present the most viable option for dealing with them.

Although analytics is used in a burgeoning array of applications (such as those found on various web sites, computing and mobile devices) it is important to note that these tools can operate independently of one another and can be deployed for specific purposes – such as for calibrating algorithms for Data Science.

On the other hand, if we go back to the Davenport's "*Competing on Analytics*" Harvard

Business Review article (2006) that kicked off the analytics movement, he defines analytics as “the ability to collect, analyze, and act on data.”

In other words, at a high level, analytics is the ability to use data to make better decisions.

Unfortunately, this does not help us much. Haven’t companies always tried to use data to make decisions?—Yes, they have. Aren’t there thousands of ways to analyze data? Yes, there are.

No wonder people are confused.

Fortunately, academic and professional organizations have realized that the field of analytics should be broken down into three categories: descriptive analytics, predictive analytics and prescriptive analytics, see Figure 12.6.

Descriptive Analytics

Descriptive analytics looks at data and analyzes past events for insight as to how to approach the future. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Almost all management reporting such as sales, marketing, operations, and finance, uses this type of post-mortem analysis.

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do.

Descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

For example, descriptive analytics examines historical electricity usage data to help plan power needs and allow electric companies to set optimal prices.

Predictive Analytics

Predictive analytics turns data into valuable, actionable information. Predictive analytics uses data to determine the probable future outcome of an event or a likelihood of a situation occurring.

Predictive analytics encompasses a variety of statistical techniques from modeling, machine learning, data mining and game theory that analyze current and historical facts to make predictions about future events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

Three basic cornerstones of predictive analytics are:

- Predictive modeling
- Decision Analysis and Optimization

- Transaction Profiling

An example of using predictive analytics is optimizing customer relationship management systems. They can help enable an organization to analyze all customer data therefore exposing patterns that predict customer behavior.

Another example is for an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

An organization must invest in a team of experts (data scientists) and create statistical algorithms for finding and accessing relevant data. The data analytics team works with business leaders to design a strategy for using predictive information.

Prescriptive Analytics

Prescriptive analytics automatically synthesizes big data, mathematical sciences, business rules, and machine learning to make predictions and then suggests decision options to take advantage of the predictions.

Prescriptive analytics goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the decision maker the implications of each decision option. Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen.

Further, prescriptive analytics can suggest decision options on how to take advantage of a future opportunity or mitigate a future risk and illustrate the implication of each decision option. In practice, prescriptive analytics can continually and automatically process new data to improve prediction accuracy and provide better decision options.

Prescriptive analytics synergistically combines data, business rules, and mathematical models. The data inputs to prescriptive analytics may come from multiple sources, internal (inside the organization) and external (social media, et al.). The data may also be structured, which includes numerical and categorical data, as well as unstructured data, such as text, images, audio, and video data, including big data. Business rules define the business process and include constraints, preferences, policies, best practices, and boundaries. Mathematical models are techniques derived from mathematical sciences and related disciplines including applied statistics, machine learning, operations research, and natural language processing.

For example, prescriptive analytics can benefit healthcare strategic planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demographic trends and population health trends, to more accurately plan for future capital investments such as new facilities and equipment utilization as well as understand the trade-offs between adding additional beds and expanding an existing facility versus building a new one.

Another example is energy and utilities. Natural gas prices fluctuate dramatically depending

upon supply, demand, econometrics, geo-politics, and weather conditions. Gas producers, transmission (pipeline) companies and utility firms have a keen interest in more accurately predicting gas prices so that they can lock in favorable terms while hedging downside risk. Prescriptive analytics can accurately predict prices by modeling internal and external variables simultaneously and also provide decision options and show the impact of each decision option.

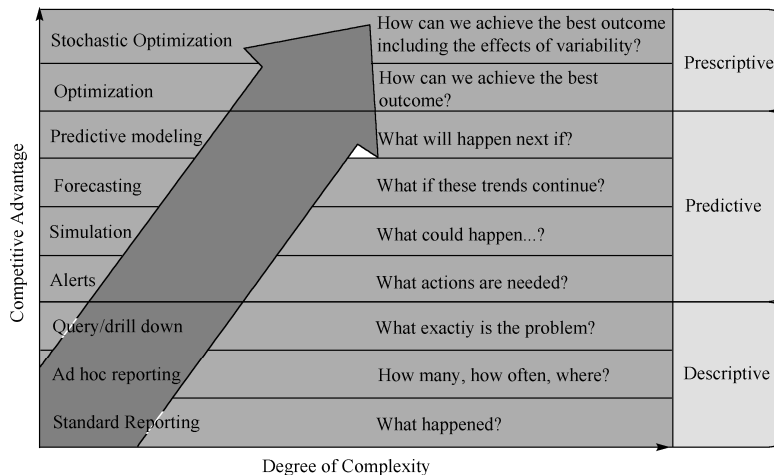


Figure 12.6 Three categories of analytics

Having this definition gives you a better framework for evaluating analytics projects. Note that this does not suggest that one type of analytics is better than another—different problems require different solutions.

When you are evaluating analytics solutions, you should understand whether the solution is descriptive, predictive, or prescriptive. Then, within each of these categories you can determine if the solution is rather basic or advanced and what will meet your needs.

12.4.2 Business Intelligence

According to Forrester, BI is:

“A set of methodologies, processes, architectures, and technologies that leverage the output of information management processes for analysis, reporting, performance management, and information delivery. Research coverage includes executive dashboards as well as query and reporting tools.”

BI is a comprehensive term that refers to analytics and reporting tools that were traditionally used to determine trends in historical data. Most vendors offer an array of tools in suites unlike analytics, which can be obtained via singular tools or applications. The key distinction between analytics and BI is that the latter actually presents the insights determined by the former in reports, dashboards, or interactive visualizations.

BI also facilitates queries in which individuals can ask data-related questions and obtain results (partly due to analytics). Unlike analytics, which is slated for those mathematically and

technologically inclined, BI tools are specifically designed to present the results of analytics in a fashion that laymen understand. The growing trend towards Data Discovery tools reinforces this capability, and helps transfer the potential of data away from IT departments and into the hands of the end user.

12.4.3 Data Science

Data Science is one of the most recent disciplines to emerge within the field of Data Management. This term is highly inclusive and was previously described by DATAVERSITY™ as:

“Data Science combines the allure of Big Data, the fascination of Unstructured Data, the precision of advanced mathematics and statistics, the innovation of social media, the creativity of storytelling, the investigation and inquiry of forensics, and the ability to use all of those skills together while still being able to demonstrate the results to non-technical audiences.”

Data Science emerged within the wake of the prevalence of Big Data. It is a term which refers to the process of deriving understanding, significance, and form from the myriads of variety of structured and unstructured Data that Big Data can encompass. Within the field, specifically trained Data Scientists create data sandboxes with which to test new forms and characteristics of data so they can ascertain what value it might have for the enterprise and how.

When organizations are utilizing different forms of data than they previously have (especially if that data is unstructured or semi-structured), Data Scientists are required to deconstruct it prior to the utilization of BI tools to gain insight from it. And, in order to successfully utilize BI and data discovery tools on such data, Data Scientists may need to develop unique algorithms both to test the data and to discern its attributes as they relate to an organization and its interests. Analytics, therefore, can play an integral role in the facilitation of this discipline.

The challenge with Data Science is all of the various skills that it requires, which expands beyond simply understanding data structure, testing and identifying it through the usage of statistics and analytics. It actually requires relating such data to an organization's objectives and being able to convey their value to IT, the business and upper level management. As such, the requirements for this science are continually varying and are shaped according to the needs of each particular enterprise.

Where first?

As the previous delineation of the distinctions between these three terms indicates, analytics is at the core of both BI and Data Science. Subsequently, there are hybrids of these terms and their technologies. Business Analytics refers to the movement of tailoring analytics and BI specifically for non-technical and business users, which typically focus on descriptive and diagnostic analytics as well as additional Data Discovery components such as search, data mashups, and geospatial technologies.

Analytics plays an integral role in the facilitation of Data Science, both during the initial

phase of testing unstructured data and while actually building applications to profit from the knowledge such data yields. Data Science is practically a requirement for Big Data initiatives (particularly those looking to leverage the wealth of unstructured and semi-structured data abounding on the Web and via the Internet of Things), yet organizations can gain simple analytic insight (even on certain forms of Big Data) by utilizing any variety of BI tools.

Finally, it is worth noting that it is also possible to use simple analytics applications, such as any one or two tools that might come in a full-fledged BI suite, to build applications to assist with data-driven business processes – either for Big Data or conventional data. This approach is something of a hybrid of all three technologies, yet also distinct in the fact that it relies more on analytics than on Data Science or merely attaining insight through BI. Application building incorporates analytics to actually create the required action that the knowledge from data provides, and is extremely organization or even business unit specific. There are a number of application building frameworks which can incorporate analytics, such as Concurrent's Cascading.

Launching Point

Although the specific approach to the application of analytics – either through BI, Data Science, or application building – may vary according to an enterprise's needs, it is important to note the broad applicability of BI. Its capacities are constantly expanding to include greater access to more forms of data in intuitive, interactive ways that favor non-technical users. Consequently, the business can do more with the data accessed through these tools in less time than it used to, which makes applying discovery-based BI an excellent starting point for the deployment of analytics. According to Gartner:

“By 2015, ‘smart data discovery,’ which includes natural-language query and search, automated, prescriptive advanced analytics and interactive data discovery capabilities, will be the most in-demand BI platform user experience paradigm, enabling mainstream business consumers to get insights (such as clusters, segments, predictions, outliers and anomalies) from data.”

New Words and Expressions

vendor ['vendə(r)] *n.* 供应商；摊贩，小贩；卖主；[贸易]自动售货机

latency ['leɪtənsɪ] *n.* 潜伏；潜在因素

boon [bu:n] *n.* 恩惠；福利 *adj.* 快乐的

empowering [ɪm'pauərɪŋ] *v.* 授权 (empower 的现在分词)；准许；增加 (某人的) 自主权

mutability [ˌmju:tə'bɪləti] *n.* 易变性；突变性；易弯性

thrive [θraɪv] *vi.* 兴盛，兴隆；长得健壮；茁壮成长

online analytical processing (OLAP) 联机分析处理；在线分析处理

burgeoning ['bɜ:dʒənɪŋ] *adj.* 迅速成长的，迅速发展的

deploy [drɪ'plɔɪ] *vt. & vi.* (尤指军事行动) 使展开；施展；有效地利用

synergistical ['saɪnədʒɪstɪkəl] *adj.* [医]协同的, 协同作用的 synergistically *adv.*
 pipeline ['paɪplaɪn] *n.* 管道; 输油管道; 渠道, 传递途径
 reinforce [ˌriːɪn'fɔːs] *vt.* 加固; 强化; 增援 *vi.* 求援; 得到增援; 给予更多的支持
 encompass [ɪn'kʌmpəs] *vt.* 围绕, 包围; 包含或包括某事物; 完成
 deconstruct [ˌdiːkən'strʌkt] *vt.* 解构 (文学作品等); 拆析
 enterprise ['entəpraɪz] *n.* 企 (事) 业单位; 事业, 计划
 delineation [dɪˌlɪn'eɪʃn] *n.* 描绘
 tailoring ['teɪləɪɪŋ] *n.* 裁缝业, 成衣业 *v.* 裁制 (tailor 的现在分词); 调整使适应
 mashup ['mæʃʌp] *n.* 混搭 (集成不同资源, 创建新的歌曲、计算机文件、程序等)
 full-fledged [fʊl fledʒd] *adj.* 经过充分训练的, 成熟的
 deployment [drɪ'plɔɪmənt] *n.* 部署; 调度
 paradigm ['pærədɑɪm] *n.* 范例, 样式, 模范

Technical Terms

1. Analytics 分析学, descriptive analytics 描述性, predictive analytics 预测性分析, prescriptive analytics 指导性分析
2. Business Intelligence, 商务智能, 商业智能, 简记为 BI
3. Gartner 高德纳, 高德纳公司成立于 1979 年(纽约交易所: IT and ITB), 是全球最具权威的 IT 市场研究与顾问咨询公司, 总部设在美国康涅狄克州斯坦福, 其研究范围覆盖全部 IT 产业, 就 IT 的研究、发展、评估、应用、市场等领域, 为客户提供客观、公正的论证报告及市场调研报告。
4. the Internet of Things 物联网
5. Concurrent 是企业大数据应用平台公司, 成立于 2008 年。该公司是流行的大数据应用开发工具 Cascading 的母公司。Cascading 可简化基于 Apache Hadoop 的大数据应用开发、部署和管理, 其月下载量超过 75000, 具体说, Cascading is a proven application development platform for building Big Data applications on Apache Hadoop.
6. Mashups 聚合是一种交互式 Web 应用程序, 它利用从外部数据源检索到的内容, 来创建全新的创新服务, 具有第二代 Web 应用程序。
7. smart data 智能数据

Notes

1. Davenport's *Competing on Analytics* 这是达文波特 (Davenport) 的一篇文章, 后来出版了同名的书, 中译本为《数据分析竞争法》(商务印书馆, 2011 年出版)。

Reading English Materials

Passage 1: The Evolution of Digital Data

Digital Data that is usually represented using binary numbers. Digital Data, in information theory and information systems, are discrete, discontinuous representations of information or works, as contrasted with continuous, or analog signals which behave in a continuous manner, or represent information using a continuous function.

Although digital representations are the subject matter of discrete mathematics, the information represented can be either discrete, such as numbers and letters, or it can be continuous, such as sounds, images, and other measurements.

The word *digital* comes from the same source as the words digit and *digitus* (the Latin word for *finger*), as fingers are often used for discrete counting. Mathematician George Stibitz of Bell Telephone Laboratories used the word *digital* in reference to the fast electric pulses emitted by a device designed to aim and fire anti-aircraft guns in 1942. The term is most commonly used in computing and electronics, especially where real-world information is converted to binary numeric form as in digital audio and digital photography.

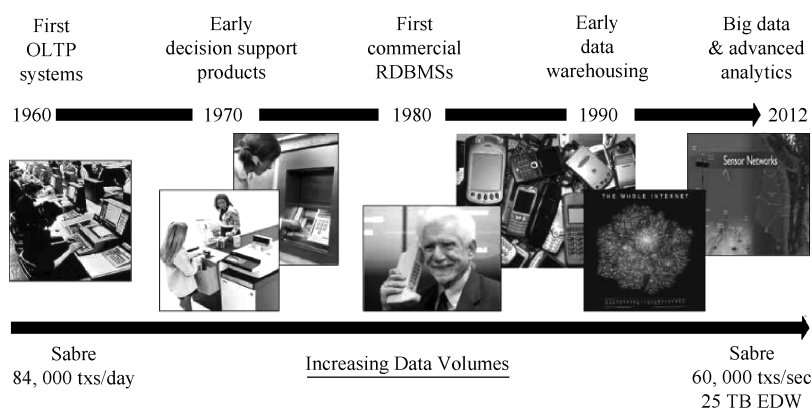


Figure 12.7 The evolution of digital data

Evolution of Digital Data Storage

Check out this cartoon I just stumbled upon on the Internet:

To those too young to understand, the first panel shows a *punch card* and the second one shows a 3.5" floppy disk. My, my, how data storage has evolved with the cloud literally taking off! See Figure 12.7.

This cartoon is from <http://awanbee.com/blog/2013/03/evolution-of-digital-data-storage/>, see Figure 12.8.

Passage 2: Big Data, Multi-Structured Data

Multi-Structured Data Definition: data that has unknown, ill-defined or overlapping schemas. Multi-structured data sets come in dozens of formats and reside in non-transactional systems such as sensors and customer interaction streams.

- Machine generated data, e. g., sensor data, system logs.
- internal/external web content including social computing data.
- Text, document and XML data.
- Graph, map and multi-media data.

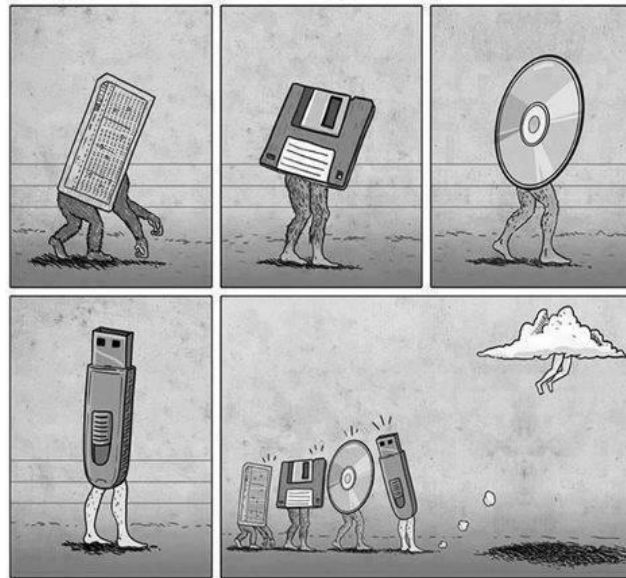


Figure 12.8 The Evolution of digital data storage

Multi-Structured Data has the following some characteristic:

- ◆ Volume increasing faster than structured data.
- ◆ Usually not integrated into a data warehouse.
- ◆ Increasing number of analytical techniques to extract useful information from this data.
- ◆ This information can be used to extend traditional predictive models and analytics.

Big data

Big data technologies apply to all types of digital data not just multi-structured data. “Big: is a relative term and is different for each organization and application.”

What you do with big data and how you use it for business benefit should be the main consideration-analytics play a key role here.

Data scientists turn into big data into big value, delivering products that delight users, and insight that informs business decisions.

Hilary Mason, *Chief Scientist at bitly think*, “A data scientist is someone who can obtain, scrub, explore, model and interpret data blending hacking, statistics and machine learning. Data scientists not only are adept at working with data, but appreciate data itself as a first-class product.”

Big data benefits have following the comparative contents, see Table 12.2.



Table 12.2 Traditional decision making environment and big data extensions

Traditional Decision-Making Environment (determine and analyze current Business situation)	Big Data Extensions (provide more complete answers, predict future business situations, Investigate new business opportunities)
Integrated data sources	Virtualized and blended data sources
Structured data	Multi-structured data
Aggregated and detailed data (with limits)	Large volumes of detailed data (no limits)
Relational EDW with at rest data Dimensional cubes/marts with at rest data	Non-relational stores with at rest data Streaming/CEP systems with in motion data
One-size fits all data management	Flexible & optimized data management
Reporting and OLAP	Advanced analytic functions & predictive models
Dashboards and scorecards	Sophisticated virtualization of large results sets
Structured navigation (drill, slice/dice)	Flexible exploration of large results sets
Humans interpret results, patterns and trends	Sophisticated trend and pattern analysis
Manual analyses, decisions and actions	Analytics & model-driven recommendations & actions

Problems

- 12.1 What is Data Science?
- 12.2 What do you think about the modern statistical analysis process?
- 12.3 What is the difference between statistician and data scientist?
- 12.4 What is statistical thinking?
- 12.5 What do you think of the two cultures of statistical modeling?

Reference

- Breiman, L., Friedman, J., Olshen, R. and Stone, C. *Classification and Regression Trees*. Wadsworth, Belmont, CA. 1984.
- Daniel, C. and Wood, F. *Fitting equations to data*. Wiley, New York. 1971.
- Davenport, T. and Patil, D.J. *Data Scientist: The Sexiest Job of the 21st Century*. 2012. <http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1>.
- Davenport, T. and Jeanne G. Harris., *Competing on Analytics: The New Science of Winning*. Harvard Business School Publishing Corporation. 2007.
- Davison, A. C. *Statistical Models*. Cambridge University Press, Cambridge. 2003.
- Vapnik, V. *The Nature of Statistical Learning Theory*. Springer, New York. 1995.
- Vapnik, V. *Statistical Learning Theory*. Wiley, New York. 1998.(中译本 ,电子工业出版社)
- Zhang, H. and Singer, B. *Recursive Partitioning in the Health Sciences*. Springer, New York. 1999.
- Wahba, G. *Spline Models for Observational Data*. SIAM, Philadelphia. 1990.
- Wickham, H., “Tidy Data,” *The Journal of Statistical Software*, 59, 1–23. 2014. Available at <http://vita.had.co.nz/papers/tidy-data.html>.
- Wickham, H., and Francois, R., *dplyr: A Grammar of Data Manipulation*. 2015. R package version 0.4.2. Available at <http://CRAN.R-project.org/package=dplyr>.
- Wilkinson, L., *The Grammar of Graphics*, New York: Springer. 2006.

Commonly Used Statistical Terms

1. 常用统计术语

Population 总体

sampling unit 抽样单元

sample 样本

observed value 观测值

descriptive statistics 描述统计学

random sample 随机样本

simple random sample 简单随机样本

statistic 统计量

order statistic 次序统计量

sample range 样本极差

mid-range 中程数

estimator 估计量 $\hat{\theta}$

sample median 样本中位数

sample moment of order k k 阶样本矩

sample mean 样本均值

average 平均数

sample variance 样本方差 S^2

sample standard deviation 样本标准差 S

sample coefficient of variation 样本变异系数

standardized sample random variable 标准化样本随机变量

sample coefficient of skewness 样本偏度系数

sample coefficient of kurtosis 样本峰度系数

sample covariance 样本协方差 S_{XY}

sample correlation coefficient 样本相关系数 r_{xy}

standard error 标准误差 $\sigma_{\hat{\theta}}$

interval estimator 区间估计量

statistical tolerance interval 统计容忍区间

statistical tolerance limit 统计容忍限

confidence interval 置信区间

one-sided confidence interval 单侧置信区间

prediction interval 预测区间

estimate 估计

error of estimation 估计误差
 bias 偏倚
 unbiased estimator 无偏估计量
 maximum likelihood estimator 极大似然估计量
 estimation 估计
 maximum likelihood estimation 极大似然估计
 maximum function 似然函数
 profile likelihood function 剖面似然函数
 hypothesis 假设： H
 null hypothesis 原假设
 alternative hypothesis 备择假设 H_A, H_1
 simple hypothesis 简单假设
 composite hypothesis 复合假设
 significance level 显著性水平
 Type I error 第一类错误
 Type II error 第二类错误
 statistical test 统计检验
 significance test 显著性检验
 p -value p 值
 power of a test 检验功效
 power curve 功效曲线
 test statistic 检验统计量
 graphical descriptive statistics 图形描述性统计量
 numerical descriptive statistics 数值描述性统计量
 classes 类, 或组
 class limits 组限
 class boundaries 组限
 mid-point of class 组中值
 class width 组距
 frequency 频数
 histogram 直方图
 bar chart 条形图
 cumulative frequency 累计频数
 relative frequency 频率
 cumulative relative frequency 累计频率

2. 概率方面术语

sample space 样本空间
 event 事件： A

complementary event 对立事件: A^c
independent event 独立事件
probability 概率
probability of an event A 事件 A 的概率
distribution function 分布函数
distribution function of a random variable X 随机变量的分布函数 $F(x)$
family of distributions 分布族
parameter 参数
random variable 随机变量
probability distribution 概率分布
expectation 期望
 p -quantile (p -fractile) 分位数 X_p, x_p
median 中位数
quartile 四分位数
univariate probability distribution 一维概率分布
univariate distribution 一维分布
multiivariate probability distribution 多维概率分布
multiivariate distribution 多维分布
marginal probability distribution 边缘概率分布
marginal distribution 边缘分布
conditional probability distribution 条件概率分布
conditional distribution 条件分布
regression curve 回归曲线
regression surface 回归曲面
discrete probability distribution 离散概率分布
discrete distribution 离散分布
continuous probability distribution 连续概率分布
continuous distribution 连续分布
probability [mass] function 概率函数
mode of probability [mass] function 概率函数的众数
probability density function 概率密度函数 $f(x)$
mode of probability density function 概率密度函数的众数
discrete random variable 离散随机变量
continous random variable 连续随机变量
centred probability distribution 中心化概率分布
centred random variable 中心化随机变量
standardized probability distribution 标准化概率分布
standardized random variable 标准化随机变量
moment of order of r r 阶[原点]矩

mean 均值 : μ
 variance 方差 : V
 standard deviation 标准差 : σ
 coefficient of variation 变异系数 : CV
 coefficient of skewness 偏度系数 : γ_1
 coefficient of kurtosis 峰度系数 : β_2
 joint moment of order r and s (r, s) 阶联合[原点]矩
 joint central moment of order r and s (r, s) 阶联合中心矩
 covariance 协方差 σ_{XY}
 correlation coefficient 相关系数
 multinomial distribution 多项分布
 binomial distribution 二项分布
 Poisson distribution 泊松分布
 hypergeometric distribution 超几何分布
 negative binomial distribution 负二项分布
 normal distribution, Gaussian distribution 正态分布
 standardized normal distribution, standardized Gaussian distribution 标准正态分布
 lognormal distribution 对数正态分布
 t distribution; Student's distribution t 分布
 degrees of freedom 自由度
 F distribution F 分布
 gamma distribution 伽玛分布
 chi-squared distribution, χ^2 distribution 卡方分布
 exponential distribution 指数分布
 beta distribution, β distribution 贝塔分布
 uniform distribution, rectangular distribution 均匀分布
 type I extreme value distribution; Gumbel distribution I 型极值分布
 type II extreme value distribution; Frechet distribution II 型极值分布
 Weibull distribution 威布尔分布
 type III extreme value distribution III 型极值分布
 multivariate normal distribution 多维正态分布
 bivariate normal distribution 二维正态分布
 standardized bivariate normal distribution 标准二维正态分布
 sampling distribution 抽样分布
 probability space 概率空间 (Ω, F, P)
 sigma algebra [of events]; σ -algebra [事件 σ] 代数
 σ -field σ 域 : F
 probability measure 概率测度

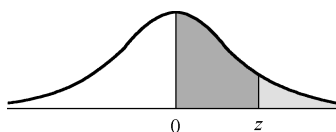
参考书：

- [1] 中国标准出版社，《中华人民共和国国家标准（GB/T 3358.1-2009/ISO 3534-1:2006 代替 GB/T 3358.1-1993）：统计学词汇及符号 第1部分：一般统计术语与用于概率的术语》。北京：2010年1月。
- [2] 中国标准出版社，《GB/T 3358.2-2009/ISO 3534-2:2006 统计学词汇及符号 第2部分：应用统计》。北京：2010年1月。
- [3] 中国标准出版社，《统计学词汇及符号（第3部分）：实验设计（GB/T 3358.3-2009/ISO 3534-3:1999）》代替 GB/T 3358.3—1993《统计学术语 第三部分 试验设计术语》。北京：2010年1月。

Appendix A Commonly Used Statistical Tables

Statistical Table 1 Areas of the Standard Normal Distribution

The entries in this table are the probabilities that a random variable with a standard normal distribution assumes a value between 0 and z ; the probability is represented by the shaded area under the curve in the accompanying figure. Areas for negative values of z are obtained by symmetry.



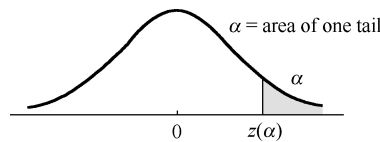
Second Decimal Place in z										
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999									
4.0	0.49997									
4.5	0.499997									
5.0	0.4999997									

For specific details about using this table to find: probabilities, see Section 6.6; confidence coefficients, see Section 6.8; p -values, Section 8.4; critical values, see Section 6.6 and Section 6.8.

Statistical Table 2 Critical Values of Standard Normal Distribution

(I) One-Tailed Situations

The entries in this table are the critical values for z for which the area under the curve representing α is in the right-hand tail. Critical values for the left-hand tail are found by symmetry. Amount of in one tail One-tailed example:

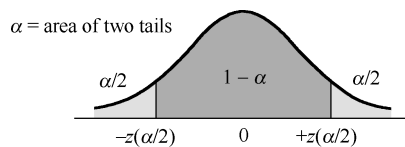


Amount of α in two-tails

α	0.25	0.10	0.05	0.025	0.02	0.01	0.005	One-tailed example: $\alpha=0.05$ $z(\alpha)=z(0.05)=1.65$
$z(\alpha)$	0.67	1.28	1.65	1.96	2.05	2.33	2.58	

(II) Two-Tailed Situations

The entries in this table are the critical values for z for which the area under the curve representing α is split equally between the two tails.



Amount of α in two-tails

α	0.25	0.20	0.10	0.05	0.02	0.01	Two-tailed example: $\alpha = 0.05$ or $1-\alpha = 0.95$ $\alpha/2 = 0.025$ $z(\alpha/2)=z(0.025)=1.96$
$z(\alpha)$	1.15	1.28	1.65	1.96	2.33	2.58	

Area in the "center"

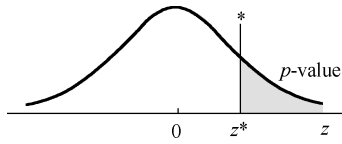
For specific details about using:

Table (I) to find: critical values, see Section 8.5.

Table (II) to find: confidence coefficients, see Section 8.2; critical values, see Section 8.2 and Section 8.5.

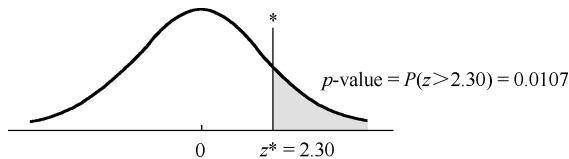
Statistical Table 3 p -Values for Standard Normal Distribution

The entries in this table are the p -values related to the right-hand tail for the calculated z^* for the standard normal distribution.



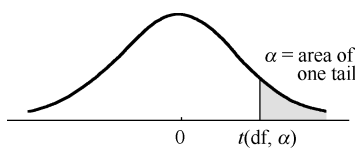
z^*	$p\text{-value}$	z^*	$p\text{-value}$	z^*	$p\text{-value}$	z^*	$p\text{-value}$
0.00	0.5000	1.00	0.1587	2.00	0.0228	3.00	0.0013
0.05	0.4801	1.05	0.1469	2.05	0.0202	3.05	0.0011
0.10	0.4602	1.10	0.1357	2.10	0.0179	3.10	0.0010
0.15	0.4404	1.15	0.1251	2.15	0.0158	3.15	0.0008
0.20	0.4207	1.20	0.1151	2.20	0.0139	3.20	0.0007
0.25	0.4013	1.25	0.1056	2.25	0.0122	3.25	0.0006
0.30	0.3821	1.30	0.0968	2.30	0.0107	3.30	0.0005
0.35	0.3632	1.35	0.0885	2.35	0.0094	3.35	0.0004
0.40	0.3446	1.40	0.0808	2.40	0.0082	3.40	0.0003
0.45	0.3264	1.45	0.0735	2.45	0.0071	3.45	0.0003
0.50	0.3085	1.50	0.0668	2.50	0.0062	3.50	0.0002
0.55	0.2912	1.55	0.0606	2.55	0.0054	3.55	0.0002
0.60	0.2743	1.60	0.0548	2.60	0.0047	3.60	0.0002
0.65	0.2578	1.65	0.0495	2.65	0.0040	3.65	0.0001
0.70	0.2420	1.70	0.0446	2.70	0.0035	3.70	0.0001
0.75	0.2266	1.75	0.0401	2.75	0.0030	3.75	0.0001
0.80	0.2119	1.80	0.0359	2.80	0.0026	3.80	0.0001
0.85	0.1977	1.85	0.0322	2.85	0.0022	3.85	0.0001
0.90	0.1841	1.90	0.0287	2.90	0.0019	3.90	0+
0.95	0.1711	1.95	0.0256	2.95	0.0016	3.95	0+

For specific details about using this table to find p -values, see Section 8.4.

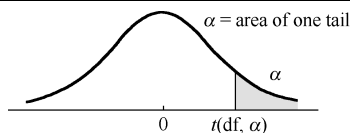


Statistical Table 4 Critical Values of Student's t -Distribution

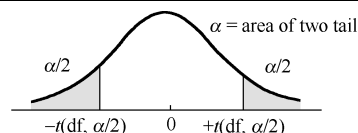
The entries in this table, $t(df, \alpha)$, are the critical values for Student's t -distribution for which the area under the curve in the right-hand tail is α . Critical values for the left-hand tail are found by symmetry.



Amount of α in One Tails						
	0.25	0.10	0.05	0.025	0.01	0.005
Amount of α in Two Tails						
df	0.50	0.20	0.10	0.05	0.02	0.01
3	0.765	1.64	2.35	3.18	4.54	5.84
4	0.741	1.53	2.13	2.78	3.75	4.60
5	0.729	1.48	2.02	2.57	3.37	4.03
6	0.718	1.44	1.94	2.45	3.14	3.71
7	0.711	1.42	1.89	2.36	3.00	3.50
8	0.706	1.40	1.86	2.31	2.90	3.36
9	0.703	1.38	1.83	2.26	2.82	3.25
10	0.700	1.37	1.81	2.23	2.76	3.17
11	0.697	1.36	1.80	2.20	2.72	3.11
12	0.696	1.36	1.78	2.18	2.68	3.05
13	0.694	1.35	1.77	2.16	2.65	3.01
14	0.692	1.35	1.76	2.14	2.62	2.98
15	0.691	1.34	1.75	2.13	2.60	2.95
16	0.690	1.34	1.75	2.12	2.58	2.92
17	0.689	1.33	1.74	2.11	2.57	2.90
18	0.688	1.33	1.73	2.10	2.55	2.88
19	0.688	1.33	1.73	2.09	2.54	2.86
20	0.687	1.33	1.72	2.09	2.53	2.85
21	0.686	1.32	1.72	2.08	2.52	2.83
22	0.686	1.32	1.72	2.07	2.51	2.82
23	0.685	1.32	1.71	2.07	2.50	2.81
24	0.685	1.32	1.71	2.06	2.49	2.80
25	0.684	1.32	1.71	2.06	2.49	2.79
26	0.684	1.32	1.71	2.06	2.48	2.78
27	0.684	1.31	1.70	2.05	2.47	2.77
28	0.683	1.31	1.70	2.05	2.47	2.76
29	0.683	1.31	1.70	2.05	2.46	2.76
30	0.683	1.31	1.70	2.04	2.46	2.75
35	0.682	1.31	1.69	2.03	2.44	2.73
40	0.681	1.30	1.68	2.02	2.42	2.70
50	0.679	1.30	1.68	2.01	2.40	2.68
70	0.678	1.29	1.67	1.99	2.38	2.65
100	0.677	1.29	1.66	1.98	2.36	2.63
df > 100	0.675	1.28	1.65	1.96	2.33	2.58



One-tailed example:
 $df = 9$ and $\alpha = 0.10$
 $t(df, \alpha) = t(9, 0.10) = 1.38$

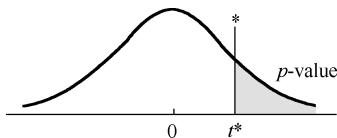


Two-tailed example:
 $df = 14$, $\alpha = 0.02$, $1 - \alpha = 0.98$
 $t(df, \alpha/2) = t(14, 0.01) = 2.62$

For specific details about using this table to find: confidence coefficients, see Section 9.1;
 p -values, pages see Section 9.1; critical values, see Section 9.1.

Statistical Table 5 Probability-Values for Student's t -distribution

The entries in this table are the p -values related to the right-hand tail for the calculated t^* value for the t -distribution of df degrees of freedom.

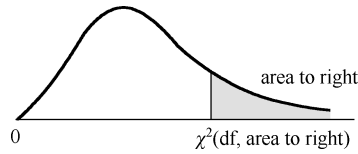


Degrees of Freedom															
t^*	3	4	5	6	7	8	10	12	15	18	21	25	29	35	$df > 45$
0.0	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500
0.1	0.463	0.463	0.462	0.462	0.462	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.461	0.460	0.460
0.2	0.427	0.426	0.425	0.424	0.424	0.423	0.423	0.422	0.422	0.422	0.422	0.422	0.421	0.421	0.421
0.3	0.392	0.390	0.388	0.387	0.386	0.386	0.385	0.385	0.384	0.384	0.384	0.383	0.383	0.383	0.383
0.4	0.358	0.355	0.353	0.352	0.351	0.350	0.349	0.348	0.347	0.347	0.347	0.346	0.346	0.346	0.346
0.5	0.326	0.322	0.319	0.317	0.316	0.315	0.314	0.313	0.312	0.312	0.311	0.311	0.310	0.310	0.310
0.6	0.295	0.290	0.287	0.285	0.284	0.283	0.281	0.280	0.279	0.278	0.277	0.277	0.277	0.276	0.276
0.7	0.267	0.261	0.258	0.255	0.253	0.252	0.250	0.249	0.247	0.246	0.246	0.245	0.245	0.244	0.244
0.8	0.241	0.234	0.230	0.227	0.225	0.223	0.221	0.220	0.218	0.217	0.216	0.216	0.215	0.215	0.214
0.9	0.217	0.210	0.205	0.201	0.199	0.197	0.195	0.193	0.191	0.190	0.189	0.188	0.188	0.187	0.186
1.0	0.196	0.187	0.182	0.178	0.175	0.173	0.170	0.169	0.167	0.165	0.164	0.163	0.163	0.162	0.161
1.1	0.176	0.167	0.161	0.157	0.154	0.152	0.149	0.146	0.144	0.143	0.142	0.141	0.140	0.139	0.139
1.2	0.158	0.148	0.142	0.138	0.135	0.132	0.129	0.127	0.124	0.123	0.122	0.121	0.120	0.119	0.118
1.3	0.142	0.132	0.125	0.121	0.117	0.115	0.111	0.109	0.107	0.105	0.104	0.103	0.102	0.101	0.100
1.4	0.128	0.117	0.110	0.106	0.102	0.100	0.096	0.093	0.091	0.089	0.088	0.087	0.086	0.085	0.084
1.5	0.115	0.104	0.097	0.092	0.089	0.086	0.082	0.080	0.077	0.075	0.074	0.073	0.072	0.071	0.070
1.6	0.104	0.092	0.085	0.080	0.077	0.074	0.070	0.068	0.065	0.064	0.062	0.061	0.060	0.059	0.058
1.7	0.094	0.082	0.075	0.070	0.066	0.064	0.060	0.057	0.055	0.053	0.052	0.051	0.050	0.049	0.048
1.8	0.085	0.073	0.066	0.061	0.057	0.055	0.051	0.049	0.046	0.044	0.043	0.042	0.041	0.040	0.039
1.9	0.077	0.065	0.058	0.053	0.050	0.047	0.043	0.041	0.038	0.037	0.036	0.035	0.034	0.033	0.032
2.0	0.070	0.058	0.051	0.046	0.043	0.040	0.037	0.034	0.032	0.030	0.029	0.028	0.027	0.027	0.026
2.1	0.063	0.052	0.045	0.040	0.037	0.034	0.031	0.029	0.027	0.025	0.024	0.023	0.022	0.022	0.021
2.2	0.058	0.046	0.040	0.035	0.032	0.029	0.026	0.024	0.022	0.021	0.020	0.019	0.018	0.017	0.016
2.3	0.052	0.041	0.035	0.031	0.027	0.025	0.022	0.020	0.018	0.017	0.016	0.015	0.014	0.014	0.013
2.4	0.048	0.037	0.031	0.027	0.024	0.022	0.019	0.017	0.015	0.014	0.013	0.012	0.012	0.011	0.010
2.5	0.044	0.033	0.027	0.023	0.020	0.018	0.016	0.014	0.012	0.011	0.010	0.010	0.009	0.009	0.008
2.6	0.040	0.030	0.024	0.020	0.018	0.016	0.013	0.012	0.010	0.009	0.008	0.008	0.007	0.007	0.006
2.7	0.037	0.027	0.021	0.018	0.015	0.014	0.011	0.010	0.008	0.007	0.007	0.006	0.006	0.005	0.005
2.8	0.034	0.024	0.019	0.016	0.013	0.012	0.009	0.008	0.007	0.006	0.005	0.005	0.005	0.004	0.004
2.9	0.031	0.022	0.017	0.014	0.011	0.010	0.008	0.007	0.005	0.005	0.004	0.004	0.004	0.003	0.003
3.0	0.029	0.020	0.015	0.012	0.010	0.009	0.007	0.006	0.004	0.004	0.003	0.003	0.003	0.002	0.002
3.1	0.027	0.018	0.013	0.011	0.009	0.007	0.006	0.005	0.004	0.003	0.003	0.002	0.002	0.002	0.002
3.2	0.025	0.016	0.012	0.009	0.008	0.006	0.005	0.004	0.003	0.002	0.002	0.002	0.002	0.001	0.001
3.3	0.023	0.015	0.011	0.008	0.007	0.005	0.004	0.003	0.002	0.002	0.002	0.001	0.001	0.001	0.001
3.4	0.021	0.014	0.010	0.007	0.006	0.005	0.003	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001
3.5	0.020	0.012	0.009	0.006	0.005	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001
3.6	0.018	0.011	0.008	0.006	0.004	0.004	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0+	0+
3.7	0.017	0.010	0.007	0.005	0.004	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0+	0+	0+
3.8	0.016	0.010	0.006	0.004	0.003	0.003	0.002	0.001	0.001	0.001	0.001	0+	0+	0+	0+
3.9	0.015	0.009	0.006	0.004	0.003	0.002	0.001	0.001	0.001	0.001	0+	0+	0+	0+	0+
4.0	0.014	0.008	0.005	0.004	0.003	0.002	0.001	0.001	0.001	0+	0+	0+	0+	0+	0+

For specific details about using this table to find p -values, see Section 9.1.

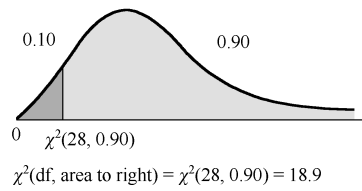
Statistical Table 6 Critical Values of χ^2 (Chi-Square) Distribution

The entries in this table, χ^2 (df, are the critical values for the χ^2 distribution for which the area under the curve to the right is α .

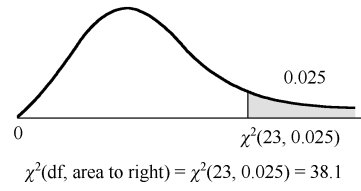


Area to the Right													
	0.995	0.99	0.975	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.025	0.01	0.005
	Area in Left-hand Tail					Median			Area in Right-hand Tail				
df	0.005	0.01	0.025	0.05	0.10	0.25	0.50	0.25	0.10	0.05	0.025	0.01	0.005
1	0.0000393	0.000157	0.000982	0.00393	0.0158	0.101	0.455	1.32	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	0.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6
3	0.0717	0.115	0.216	0.352	0.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8
4	0.207	0.297	0.484	0.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9
5	0.412	0.554	0.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.8
6	0.676	0.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.5	16.8	18.6
7	0.990	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.7	17.3	19.7	21.9	24.7	26.8
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.8	18.5	21.0	23.3	26.2	28.3
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	16.0	19.8	22.4	24.7	27.7	29.8
14	4.07	4.66	5.63	6.57	7.79	10.2	13.34	17.1	21.1	23.7	26.1	29.1	31.3
15	4.60	5.23	6.26	7.26	8.55	11.0	14.34	18.2	22.3	25.0	27.5	30.6	32.8
16	5.14	5.81	6.91	7.96	9.31	11.9	15.34	19.4	23.5	26.3	28.8	32.0	34.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.34	20.5	24.8	27.6	30.2	33.4	35.7
18	6.26	7.01	8.23	9.39	10.9	13.7	17.34	21.6	26.0	28.9	31.5	34.8	37.2
19	6.84	7.63	8.91	10.1	11.7	14.6	18.34	22.7	27.2	30.1	32.9	36.2	38.6
20	7.43	8.26	9.59	10.9	12.4	15.5	19.34	23.8	28.4	31.4	34.2	37.6	40.0
21	8.03	8.90	10.3	11.6	13.2	16.3	20.34	24.9	29.6	32.7	35.5	38.9	41.4
22	8.64	9.54	11.0	12.3	14.0	17.2	21.34	26.0	30.8	33.9	36.8	40.3	42.8
23	9.26	10.2	11.7	13.1	14.8	18.1	22.34	27.1	32.0	35.2	38.1	41.6	44.2
24	9.89	10.9	12.4	13.8	15.7	19.0	23.34	28.2	33.2	36.4	39.4	43.0	45.6
25	10.5	11.5	13.1	14.6	16.5	19.9	24.34	29.3	34.4	37.7	40.6	44.3	46.9
26	11.2	12.2	13.8	15.4	17.3	20.8	25.34	30.4	35.6	38.9	41.9	45.6	48.3
27	11.8	12.9	14.6	16.2	18.1	21.7	26.34	31.5	36.7	40.1	43.2	47.0	49.6
28	12.5	13.6	15.3	16.9	18.9	22.7	27.34	32.6	37.9	41.3	44.5	48.3	51.0
29	13.1	14.3	16.0	17.7	19.8	23.6	28.34	33.7	39.1	42.6	45.7	49.6	52.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.34	34.8	40.3	43.8	47.0	50.9	53.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.34	45.6	51.8	55.8	59.3	63.7	66.8
50	28.0	29.7	32.4	34.8	37.7	42.9	49.33	56.3	63.2	67.5	71.4	76.2	79.5
60	35.5	37.5	40.5	43.2	46.5	52.3	59.33	67.0	74.4	79.1	83.3	88.4	92.0
70	43.3	45.4	48.8	51.7	55.3	61.7	69.33	77.6	85.5	90.5	95.0	100.4	104.2
80	51.2	53.5	57.2	60.4	64.3	71.1	79.33	88.1	96.6	101.9	106.6	112.3	116.3
90	59.2	61.8	65.6	69.1	73.3	80.6	89.33	98.6	107.6	113.1	118.1	124.1	128.3
100	67.3	70.1	74.2	77.9	82.4	90.1	99.33	109.1	118.5	124.3	129.6	135.8	140.2

Left-tail example:
Find χ^2 with $df = 28$; area in left-tail = 0.10



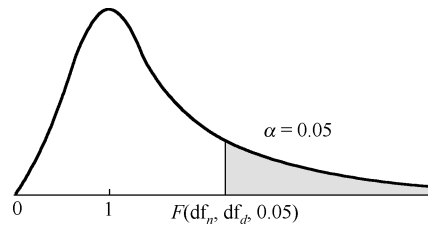
Right-tail example:
Find χ^2 with $df = 23$; area in right-tail = 0.025



For specific details about using this table to find: p -values, see Section 9.3; critical values, see Section 9.3.

Statistical Table 7(I) Critical Values of the F Distribution ($\alpha = 0.05$)

The entries in this table are critical values of F for which the area under the curve to the right is equal to 0.05.



		Degrees of Freedom for Numerator									
		1	2	3	4	5	6	7	8	9	10
Degrees of Freedom for Denominator	1	161.	200.	216.	225.	230.	234.	237.	239.	241.	242.
	2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4
	3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
	12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
	13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
	25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
	60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
	120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
		3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

For specific details about using this table to find: p -values, see Section 10.5.2; critical values, see Section 10.5.2.

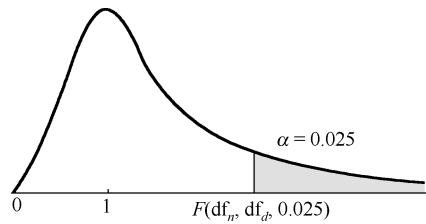
Table 7(I) (Continued)

		Degrees of Freedom for Numerator								
		12	15	20	24	30	40	60	120	
Degrees of Freedom for Denominator	1	244.	246.	248.	249.	250.	251.	252.	253.	2.54.
	2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5
	3	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
	4	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
	5	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
	6	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
	7	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
	8	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
	9	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
	10	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
	11	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
	12	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
	13	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
	14	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
	15	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
	16	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
	17	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
	18	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
	19	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
	20	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
	21	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
	22	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
	23	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
	24	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
	25	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
	30	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
	40	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
	60	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
	120	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
		1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

From E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1 (1958), pp. 159-163. Reprinted by permission of the Biometrika Trustees.

Statistical Table 7(II) Critical Values of the F Distribution ($\alpha = 0.025$)

The entries in this table are critical values of F for which the area under the curve to the right is equal to 0.025.



		Degrees of Freedom for Numerator									
		1	2	3	4	5	6	7	8	9	10
Degrees of Freedom for Denominator	1	648.	800.	864.	900.	922.	937.	948.	957.	963.	969.
	2	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4	39.4
	3	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5	14.4
	4	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84
	5	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62
	6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46
	7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76
	8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30
	9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96
	10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72
	11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53
	12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37
	13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25
	14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.28	3.21	3.15
	15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06
	16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99
	17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92
	18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87
	19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82
	20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77
	21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73
	22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70
	23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67
	24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64
	25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61
	30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51
	40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39
	60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27
	120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16
		5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05

For specific details about using this table to find: p -values, see Section 10.5.2; critical values, see Section 10.5.2.

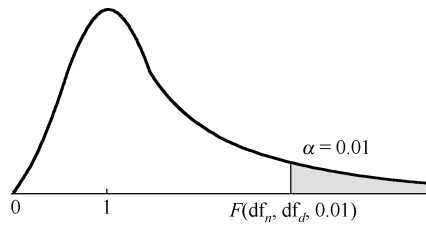
Table 7(II) (Continued)

		Degrees of Freedom for Numerator								
		12	15	20	24	30	40	60	120	
Degrees of Freedom for Denominator	1	977.	985.	993.	997.	1001.	1006.	1010.	1014.	1018.
	2	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5	39.5
	3	14.3	14.3	14.2	14.1	14.1	14.0	14.0	13.9	13.9
	4	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
	5	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
	6	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
	7	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
	8	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
	9	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
	10	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
	11	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
	12	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
	13	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
	14	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
	15	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
	16	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
	17	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
	18	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
	19	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
	20	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
	21	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
	22	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
	23	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
	24	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
	25	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
	30	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
	40	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
	60	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
	120	2.05	1.95	1.82	1.76	1.69	1.61	1.53	1.43	1.31
		1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

From E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1 (1958), pp. 159-163. Reprinted by permission of the Biometrika Trustees.

Statistical Table 7(III) Critical Values Of the F Distribution ($\alpha = 0.01$)

The entries in the table are critical values of F for which the area under the curve to the right is equal to 0.01



		Degrees of Freedom for Numerator									
		1	2	3	4	5	6	7	8	9	10
Degrees of Freedom for Denominator	1	4052.	5000.	5403.	5625.	5764.	5859.	5928.	5982.	6024.	6056.
	2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4
	3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2
	4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5
	5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1
	6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
	7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
	8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
	9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
	10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
	11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
	12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
	13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
	14	8.86	6.51	5.56	5.04	4.70	4.46	4.28	4.14	4.03	3.94
	15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
	16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
	17	8.40	6.11	5.19	4.67	4.34	4.10	3.93	3.79	3.68	3.59
	18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
	19	8.19	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43
	20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
	21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
	22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
	23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
	24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
	25	7.77	5.57	4.68	4.18	3.86	3.63	3.46	3.32	3.22	3.13
	30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
	40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80
	60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
	120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47
		6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32

For specific details about using this table to find: p -values, see Section 10.5.2; critical values, see Section 10.5.2.

Table 7(III) (Continued)

		Degrees of Freedom for Numerator								
		12	15	20	24	30	40	60	120	
Degrees of Freedom for Denominator	1	6106.	6157.	6209.	6235.	6261.	6287.	6313	6339.	6366.
	2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5
	3	27.1	26.9	26.7	26.6	26.5	26.4	26.3	26.2	26.1
	4	14.4	14.2	14.0	13.9	13.8	13.7	13.7	13.6	13.5
	5	9.89	9.72	9.55	9.47	9.38	9.29	9.20	9.11	9.02
	6	7.72	7.56	7.40	7.31	7.23	7.14	7.06	6.97	6.88
	7	6.47	6.31	6.16	6.07	5.99	5.91	5.82	5.74	5.65
	8	5.67	5.52	5.36	5.28	5.20	5.12	5.03	4.95	4.86
	9	5.11	4.96	4.81	4.73	4.65	4.57	4.48	4.40	4.31
	10	4.71	4.56	4.41	4.33	4.25	4.17	4.08	4.00	3.91
	11	4.40	4.25	4.10	4.02	3.94	3.86	3.78	3.69	3.60
	12	4.16	4.01	3.86	3.78	3.70	3.62	3.54	3.45	3.36
	13	3.96	3.82	3.66	3.59	3.51	3.43	3.34	3.25	3.17
	14	3.80	3.66	3.51	3.43	3.35	3.27	3.18	3.09	3.00
	15	3.67	3.52	3.37	3.29	3.21	3.13	3.05	2.96	2.87
	16	3.55	3.41	3.26	3.18	3.10	3.02	2.93	2.84	2.75
	17	3.46	3.31	3.16	3.08	3.00	2.92	2.83	2.75	2.65
	18	3.37	3.23	3.08	3.00	2.92	2.84	2.75	2.66	2.57
	19	3.30	3.15	3.00	2.92	2.84	2.76	2.67	2.58	2.49
	20	3.23	3.09	2.94	2.86	2.78	2.69	2.61	2.52	2.42
	21	3.17	3.03	2.88	2.80	2.72	2.64	2.55	2.46	2.36
	22	3.12	2.98	2.83	2.75	2.67	2.58	2.50	2.40	2.31
	23	3.07	2.93	2.78	2.70	2.62	2.54	2.45	2.35	2.26
	24	3.03	2.89	2.74	2.66	2.58	2.49	2.40	2.31	2.21
	25	2.99	2.85	2.70	2.62	2.53	2.45	2.36	2.27	2.17
	30	2.84	2.70	2.55	2.47	2.39	2.30	2.21	2.11	2.01
	40	2.66	2.52	2.37	2.29	2.20	2.11	2.02	1.92	1.80
	60	2.50	2.35	2.20	2.12	2.03	1.94	1.84	1.73	1.60
	120	2.34	2.19	2.03	1.95	1.86	1.76	1.66	1.53	1.38
		2.18	2.04	1.88	1.79	1.70	1.59	1.47	1.32	1.00

From E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, vol. 1 (1958), pp. 159-163. Reprinted by permission of the Biometrika Trustees.

Appendix B Summary of Univariate Descriptive Statistics and Graphs for the Four Level of Measurement

B.1 Level of measurement: Properties, Response Formats, Descriptive statistics, Graphs

Level of measurement	Properties	Response Formats and Examples	Descriptive statistics	Graphs
Nominal/ Categorical	Discrete (D) Arbitrary (no order)	Dichotomous-Gender, Multichotomous-Religion	Mode Frequencies/Percentages	Bar/Pie
Ordinal / Rank	Discrete Ordered /Ranked	Ranking-Grade (F, P, CR, DI, HD)	Mode Frequencies/Percentages Mix/Max/Range Median/Percentiles (if meaningful)	Bar/Pie
Interval	Equal distance between values 0 is arbitrary More than approx. 5 values can be treated as continuous	Likert scale—Attitude Semantic differential Composite scores	Mode (D) Frequencies/Percentages(D) Mix/Max/Range Median/Percentiles Mean/SD/Skew/Kurt	Bar/Pie Stem & Leaf (if D/rounded) Boxplot/Error-bar Histogram (if Metric)
Ratio	Meaningful 0 Continuous/Metric (M) Ratio-type observations can be made	Numeric — Age, Height, Weight, Number of times an event occurs	Mode/Frequencies/Percentages (if meaningful) Mix/Max/Range Median/Percentiles Mean/SD/Skew/Kurt	Histogram Stem & Leaf (if D/rounded) Boxplot/Error-bar

B.2 Summary of Descriptive Statistics & Graphical Summaries for the Four Levels of Measurement

Statistic	Nominal	Ordinal	Interval	Ratio
Mode	√	√	√	If meaningful
Median	×	√	√	√
Range, Min. Max	×	√	√	√
Mean	×	×	If metric	√
SD	×	×	If metric	√

B.3 Summary of Descriptive Statistics & Graphical Summaries for the Four Levels of Measurement

Graph	Nominal	Ordinal	Interval	Ratio
Bar / Pie	√	√	If discrete	×
Stem & Leaf	×	√	√	√
Boxplot	×	√	√	√
Histogram	×	×	If metric	√

Appendix C Order of Magnitude of Data

C.1 Order of Magnitude of Data (数据阶常用英文表示法)

In words	Value	Power of ten	Order of magnitude	Prefix	Metric	Abbreviation	中文
one	1	10^0	0	—	—	—	—
ten	10	10^1	1	deca-	—	—	—
hundred	100	10^2	2	hecto-	—	—	—
thousand	1,000	10^3	3	kilo-	kilobytes	kB	千字节
million	$1,000^2$	10^6	6	mega-	megabyte	MB	兆字节
billion	$1,000^3$	10^9	9	giga-	gigabyte	GB	吉字节
trillion	$1,000^4$	10^{12}	12	tera-	terabyte	TB	太字节
quadrillion	$1,000^5$	10^{15}	15	peta-	petabyte	PB	拍字节
quintillion	$1,000^6$	10^{18}	18	exa-	exabyte	EB	艾字节
sextillion	$1,000^7$	10^{21}	21	zetta-	zettabyte	ZB	泽字节
septillion	$1,000^8$	10^{24}	24	yotta-	yottabyte	YB	尧字节

注释：(1) bit：比特，二进制的位，存放一位二进制数，即 0 或 1。(2) byte：字节、字组作为一个单位来操作（运算）的二进制字符串，通常 8bits 组成一个字节，也就是 1byte = 8bits，byte 简写为 B。

C.2 Data Representation in Computer Systems Storage Unit

(计算机系统存储单位数据表示)

单位名称	换算等式	形象注释
Bit (位)	存放一位二进制数，即 0 或 1	最小的存储单位
Byte (字节)记为 B	8 个二进制为一个字节(B)	常用的单位，一个汉字占 2 个字节
KB (千字节)	1KB = 1024 B	1 千个字节约等于 512 个汉字的存储
MB (兆字节)	1MB = 1024 KB	1 兆字节约等于存储 52 万多个汉字，相当于存储一本 50 多万字的书
GB (吉字节)	1GB = 1024 MB	1GB 相当于存储 1 千多本 50 多万字的书
TB (太字节)	1TB = 1024 GB	1TB 相当于存储 100 万本 50 多万字的书
PB (拍字节)	1PB = 1024 TB	1PB 相当于存储 10 亿多本 50 多万字的书
EB (艾字节)	1EB = 1024 PB	1EB 相当于存储全球所有图书
ZB (泽字节)	1ZB = 1024 EB	1ZB 相当于存储全球目前所有互联网数据信息
YB (尧字节)	1YB = 1024 ZB	1YB 相当于存储全球人类所有数据信息。人类目前的所有互联网数据、移动数据、书籍数据加起来都不够 1YB 的存储容量

注释：bits short for binary digits.

References

- [1] Agresti, Alan and Finlay, Barbara. , *Statistical Methods for Social Sciences. Fourth edition*, Pearson Education Asia Ltd, and 2009.
- [2] Breiman, Loe. , Statistical Modeling: The Two Cultures. *Statistical Science*, Vol 16, No 3, 199-231. 2001.
- [3] Cameron, Colin, A. and Pravin K. Trivedi., *Microeconometrics: Methods and Applications*, Cambridge University Press. 2005.
- [4] Casella. George, and Berger, Roger L. Statistical Inference. 2nd Edition. Cengage Learning. 2001.
- [5] Davenport, Thomas H. and Harris, Jeanne G. *Competing on Analytics: The New Science of Winning*. Harvard Business Review Press. 2007.
- [6] Davenport, Thomas H. and D.J. Patil. *Data Scientist: The Sexiest Job of the 21st Century*. Harvard Business Review. September. 2012.
- [7] Foster Provost. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media. 2013.
- [8] Gilchrist, W. G., *Statistical Modelling with Quantile Functions*, Chapman & Hall / CRC, London, England. 2000.
- [9] Hey, Tony. and Tolle, Kristin. *The Fourth Paradigm: Data –Intensive Scientific Discovery*, Microsoft research. 2010.
- [10] Efraim Turban, Jay E. Aronson, Ting-Peng Liang, Ramesh Sharda., *Decision Support and Business Intelligence Systems*. Pearson Pretince Hall, New Jersey, 2007.
- [11] Hogg, R.V. and McKean, J. W., Craig, A.T., *Introduction to Mathematical Statistics* (Seventh Edition), Prentice Hall. 2013.
- [12] Michael Minelli, Michele Chambers, Ambiga Dhiraj., *Big Data, Big Analytics: Emerging Business Intelligence and Analytic Trends for Today's Businesses*, Wiley, 2013.
- [13] Paul D. Velleman, Noreen D. Sharpe and Richard D. De Veaux., *Business Statistics*. 3rd. Pearson. 2014.
- [14] Savage. Sam., and Danziger. Jeff, *The Flaw of Averages: Why We Underestimate Risk in the Face of Uncertainty*. Wiley. 2012.
- [15] Tabak, John. *Probability and Statistics: The Science of Uncertainty*, Checkmark Books. 2005.
- [16] Vapnik, Vladimir N. *Statistical Learning Theory*. John Wiley & Sons, Inc. 1998.
- [17] Vapnik, Vladimir N. *The Nature of Statistical Learning Theory*. Springer. 1998.
- [18] 陆谷孙, 《英汉大词典(第2版)》, 上海译文出版社, 2007.
- [19] 张鸿林, 葛显良, 《英汉数学词汇(第2版)》, 清华大学出版社, 2010.
- [20] 王忠玉, 《统计学专业英语(第3版)》, 哈尔滨工业大学出版社, 2015.